

**CRESST REPORT 750**

*Julia Phelan  
Taehoon Kang  
David N. Niemi  
Terry Vendlinski  
Kilchan Choi*

**SOME ASPECTS OF THE  
TECHNICAL QUALITY OF  
FORMATIVE ASSESSMENTS  
IN MIDDLE SCHOOL  
MATHEMATICS**

JANUARY, 2009



**National Center for Research on Evaluation, Standards, and Student Testing**

Graduate School of Education & Information Studies  
UCLA | University of California, Los Angeles



**Some Aspects of the Technical Quality of Formative Assessments  
in Middle School Mathematics**

CRESST Report 750

Julia Phelan, Taehoon Kang, David N. Niemi, Terry Vendlinski & Kilchan Choi  
CRESST/University of California, Los Angeles

January, 2009

National Center for Research on Evaluation,  
Standards, and Student Testing (CRESST)  
Center for the Study of Evaluation (CSE)  
Graduate School of Education & Information Studies  
University of California, Los Angeles  
300 Charles E. Young Drive North  
GSE&IS Building, Box 951522  
Los Angeles, CA 90095-1522  
(310) 206-1532

Copyright © 2009 The Regents of the University of California

The work reported herein was supported under the National Research and Development Centers, Award Number R305A050004, as administered by the U.S. Department of Education's Institute of Education Sciences (IES).

The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the National Research and Development Centers or the U.S. Department of Education's Institute of Education Sciences (IES).

# **SOME ASPECTS OF THE TECHNICAL QUALITY OF FORMATIVE ASSESSMENTS IN MIDDLE SCHOOL MATHEMATICS**

Julia Phelan, Taehoon Kang, David N. Niemi,  
Terry Vendlinski and Kilchan Choi  
CRESST/University of California, Los Angeles

## **Abstract**

While research suggests that formative assessment can be a powerful tool to support teaching and learning, efforts to jump on the formative assessment bandwagon have been more widespread than those to assure the technical quality of the assessments. This report covers initial analyses of data bearing on the quality of formative assessments in middle school mathematics. Specifically, these data address the question of whether relatively short assessments can provide reliable and useful information on middle school students' understanding of conceptual domains in pre-algebra. Items and test forms were developed and tested in four domains (rational number equivalence, properties of arithmetic, principles for solving equations, and applications of these concepts to other domains), all of which are critical to eventual mastery of algebra. We tested the items with sixth-grade students in classrooms in four districts. We then pared down the items to create eight assessment forms that were further tested alongside instructional support materials and professional development. Results of this study suggest that relatively brief formative assessments focused on key conceptual domains can provide reliable and useful information on students' levels of understanding and possible misunderstandings in the domain.

## **Introduction**

Improving formative assessment has been widely touted as means to improving student learning—and as a possible counterweight to the narrowing of the curriculum effected by state testing programs. Although researchers have demonstrated that formative assessment can effectively improve student achievement and understanding (see, Black & Wiliam, 1998a, 1998b; Black, Harrison, Lee, Marshall, & Wiliam, 2003; Minstrell, 2000; Niemi, 1996; Pellegrino, Chudowsky, & Glaser, 2001) the focus on assuring the technical quality of these assessments has been less widespread. As the AERA, NCME, APA (1999) *Standards for educational and psychological testing* advocate, all assessments should be validated for their intended purposes, and in the case of formative assessments, this means that evidence should be obtained to show that the assessments provide information to effectively guide instruction, that teachers use that information to change the course of instruction (when appropriate), and that learning is ultimately improved.

Not only is there scant evidence on the validity of district benchmark or classroom assessments intended for formative purposes, but anecdotal and research evidence from districts across the U.S. suggests that many teachers are unable to use the information from benchmark tests or their own assessments because they lack the knowledge, materials, or curricular time to do so. As a result, there is a great deal of rhetoric surrounding formative assessment and “doing something” with the results, but in reality teachers don’t always have the wherewithal to do anything except repeat what they have already done. As Stiggins (2004) notes, “teacher must possess and be ready to apply knowledge of sound classroom assessment practices...if teachers assess accurately and use the results effectively, then students will prosper,” (p. 26).

### **Description of POWERSOURCE<sup>®</sup> and its Rationale**

The study we describe in this report is embedded in a larger formative assessment project that seeks to help assure that students possess key understandings they need for success in Algebra I. Such an emphasis is strategic because failure to master Algebra I keeps many students from advancing in mathematics. Indeed there is ample evidence showing the frequency and price of failure for subsequent academic performance, including high school graduation, college entry and preparation (e. g., Brown & Niemi, 2007). For example, data from the California State Algebra I exam over the past 5 years reveals that on average, 76% of students are below proficiency (California Department of Education, 2007).

The assessments we tested in this study are one component of an intervention that includes professional development, instructional activities, and resources to help students who have not mastered the big ideas. This comprehensive intervention—called POWERSOURCE<sup>®</sup>—is currently being experimentally tested in the 2007–08 school year in a large sample of classrooms. Using longitudinal designs (students will participate in both their sixth- and seventh-grade years, and their achievement will be measured multiple times in both years) and hierarchical linear modeling, our studies will examine research questions dealing with the differential effects of POWERSOURCE<sup>®</sup> on student learning (as measured by state and district tests, our own posttests, and transfer tasks drawn from international tests), teachers’ effectiveness, and moderating variables that influence success. We will also be collecting extensive data on the accuracy and reliability of the information provided by the assessments, and the relationships between the assessments, large-scale accountability tests, and transfer measures. We expect the research to produce important scientific knowledge as well as validated model formative assessments, instructional tools, and professional development for practitioners and policy guidance. In this report, however, we concentrate on our research to establish the quality of the POWERSOURCE<sup>®</sup> measures.

Development of the assessments was preceded by a detailed analysis of the domains to be assessed and the creation of an algebra ontology used to help create the assessment items (Niemi & Phelan, 2000; Niemi, Vallone, & Vendlinski, 2006). First we worked with mathematicians and mathematics educators to develop an ontology or conceptual map of algebra knowledge and its prerequisites. Rather than consider how topics are typically organized for pedagogical purposes, or how content is conventionally organized in curricula, the experts were asked to consider which ideas were most important in their own thinking and problem solving. The expert panel first identified the big ideas that organized their thinking and work in the domain, then subordinate or supporting ideas that elaborated and gave meaning to the organizing concepts. The list of big ideas and supporting ideas was subsequently reviewed and the language slightly revised by six other mathematicians working in two separate groups (Niemi & Phelan, 2000). An ontology showing the big ideas and relationships between them is shown in Figure 1 (for the final list of big ideas see Appendix A).

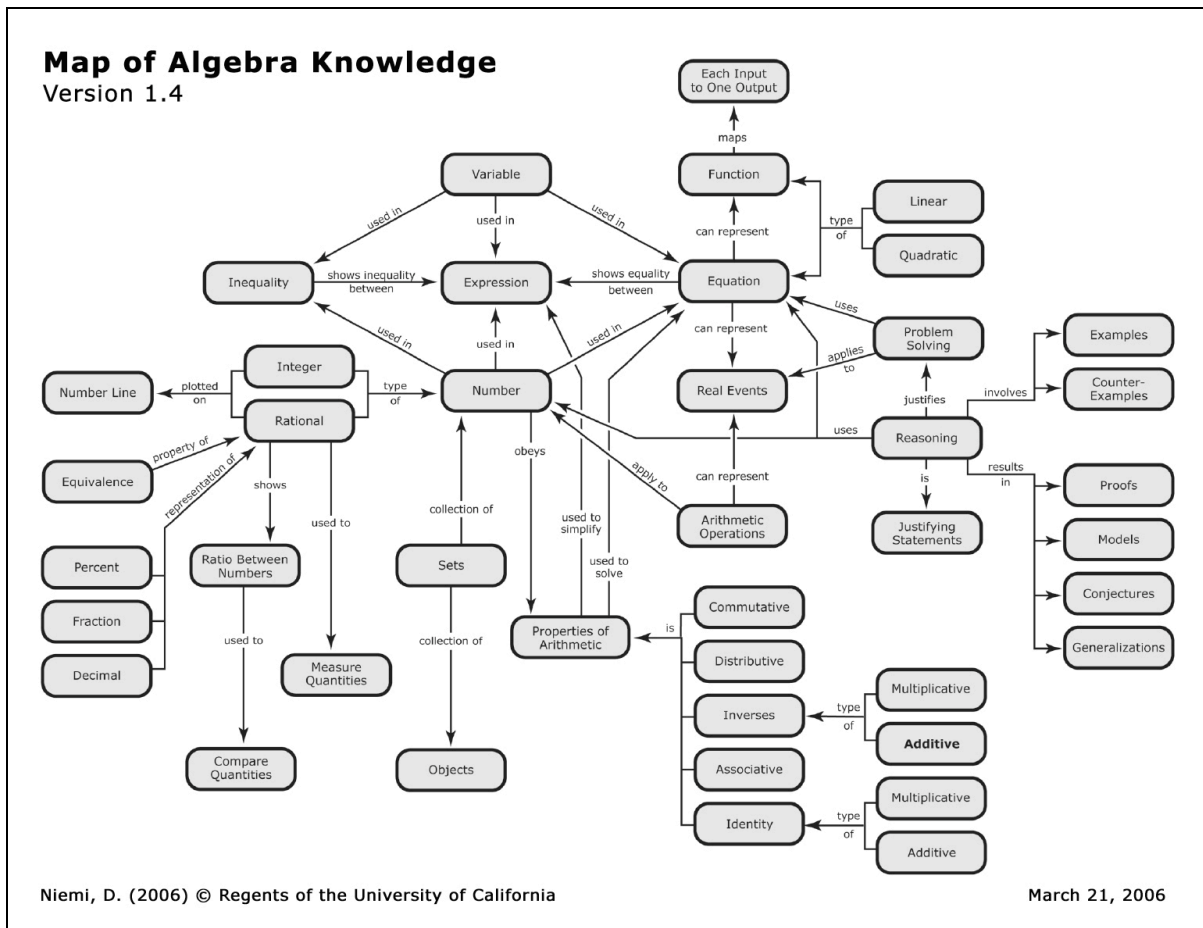


Figure 1. Map showing relationships of big ideas.

We used this ontology to develop and test a series of diagnostic assessments that teachers can use in elementary through middle school to help assure that their students are developing the fundamental knowledge and skills they will need to succeed in algebra. Central to POWERSOURCE<sup>®</sup> is providing teachers with materials and concrete guidance on how to use student performance data to improve instruction, thus directly impacting teacher instructional capacity.

In this report we focus on the development and pilot testing of formative assessments used in the POWERSOURCE<sup>®</sup> sequence. Our goal was to obtain information that would help us to refine the assessments prior to using them as formative assessments in a larger-scale study. The major steps in developing and pilot testing items are as follows:

1. Select conceptual domains to be assessed using the formative assessments.
2. Develop a large set of items covering our selected domains at the sixth-grade level.
3. Group these items into assessment forms called *Checks for Understanding*.
4. Pilot test these forms in sixth-grade classrooms.
5. Analyze pilot test data and use the results of the pilot studies to refine the item set and select items for additional field testing alongside instructional resources and professional development.

These steps are described in more detail below.

**1) Selecting the domains:** Because formative assessment can be costly and time-consuming to implement, we decided to focus the POWERSOURCE<sup>®</sup> intervention on some of the most important conceptual domains in pre-algebra. Four domains were chosen because they are widely considered to essential to later mastery of algebra and their significant place in state mathematics standards across Grades 6–8:

- ∞ Properties of arithmetic, particularly the distributive property and its use in transforming expressions to solve equations.
- ∞ Principles for solving equations, including knowledge of what it means to solve an equation, the meaning of the equals sign, principles underlying solution procedures for simple linear equations, and principles for writing equations to represent real situations.
- ∞ Rational number equivalence: The big idea here is that equivalent representations of a rational number can be generated by applying the multiplicative identity (e.g., multiply a given fraction by  $a/a$ , where  $a \neq 0$ ). Examples of related procedural knowledge include procedures for recognizing and finding equivalent representations of fractions, finding fractions that are equivalent to a given fraction; and determining when two fractions are equivalent.



∞Application of concepts and principles: In addition to these three domains, we also investigated their potential applications to other domains; for example, how the principles of solving equations could be used to solve geometry problems.

**2) Item Development:** Item development procedures drew on assessment models we have validated in an extensive series of studies over many years (e.g., Baker, Freeman, & Clayton, 1991; Niemi, 1996; Baker, 1997; Niemi, Baker, & Sylvester, 2007). In each domain we developed item architectures covering a range of cognitive demands, including use of mathematical representations, computation skills, problem solving, and understanding and explanation of key concepts. Assessment development teams composed of mathematics educators and experienced item writers developed an initial pool of items; then mathematics educators and project staff reviewed these items for content relevance and potential bias with then the items passing this review were assembled into forms for further testing.

### **Item Types**

We developed and tested seven different item architectures to see which ones, or which combinations would give us the most useful information and perform most reliably in combination with each other. These item types are: (a) basic computation and symbolic representation tasks, (b–c) partially worked problems (with or without explanations), (d) explanation tasks, (e) word problems, (f) other complex problems involving graphics, and (g) tasks embedded in narratives.

### **Basic Computation/Representation**

The basic computation/representation tasks were designed to assess whether or not students can recognize problems as instances of particular ideas and can then solve the tasks successfully. Tasks are simple, well-defined problems representing an application of the relevant big idea, (e.g. for the distributive property a typical task is):

$$6(3 + 1) = 6 \times \square + 6 \times 1$$

With respect to the second category of tasks, partially-worked problems, some evidence suggests that learning from worked examples of problem solutions is an effective way to develop cognitive skills in well structured domains such as math and physics (Renkl, Atkinson, Maier, 2000; Sweller, van Merriënboer, & Paas, 1998; VanLehn, 1996). Indeed, Sweller et al. (1989) found that studying worked examples, then gradually learning to solve partly-solved problems, and finally complete problems with no scaffolded help, was more effective than traditional problem-solving instruction alone. We decided to investigate whether partially-worked examples, particularly those requiring explanations, may allow us

to make inferences about students' understanding of problem solving procedures rather than merely providing information about the ability to recall and execute procedures.

In the partially-worked examples that we have tested as possible assessments, the student must read and understand problem-solving steps completed by another person and fill in one to three boxes representing missing numbers or symbols in the problem solution, or fill in a complete problem solving step (see Figure 2 for an example). These problems are preceded by a short fully-worked example (Figure 3), involving no more than 3–4 problem solving steps, which the student can read to see what is expected. The fully-worked example covers a topic similar, but not identical to the topic to be assessed.

Problem Solving Step	
1	$x + 5 = 10$
2	$x + 5 - 5 = 10 - \square$
3	$x = \square$

Simplifying Step	
1	$2(7 + 4)$
2	$2 \cdot 7 + 2 \cdot 4$
3	<div style="border: 1px solid black; height: 20px; width: 100%;"></div>
4	22

Figure 2. Examples of partially-worked problems.

	Simplifying Step	Explanation
1	$(5 - 3)^2 - 2$	Write the problem.
2	$(2)^2 - 2$	Subtract the numbers in the parenthesis because that comes first in the order of operations.
3	$4 - 2$	Find the value of the exponents because you calculate exponents next in the order of operations.
4	2	Subtract to get the answer.

Figure 3. A fully-worked example.

The third type of assessment developed is the partially-worked example with justifications. In order to solve these problems, the student must read and understand problem-solving steps completed by another person, and must provide a principled explanation for one of these steps (see Figure 4). As in the case of partially-worked

problems, these problems are preceded by a short fully-worked example containing no more than 3–4 problem-solving steps and covering a related but not identical topic.

④ A student solved the problem  $\frac{7}{12} + \frac{1}{4}$  the following way. Explain why the student multiplied  $\frac{1}{4}$  by  $\frac{3}{3}$  in step 2. *Be sure to use some mathematical rule or principle in your explanation.*

	Problem Solving Step	Explanation
1	$\frac{7}{12} + \frac{1}{4} =$	These fractions have different denominators. To add these fractions, I need to find a common denominator.
2	$\frac{7}{12} + \left( \frac{1}{4} \times \frac{3}{3} \right)$	<p>a) Why did the student multiply <math>\frac{1}{4}</math> by <math>\frac{3}{3}</math> ?</p> <div style="border: 1px solid black; height: 80px; margin: 10px 0;"></div> <p>b) Is it mathematically ok to multiply <math>\frac{1}{4}</math> by <math>\frac{3}{3}</math> ? Explain why or why not.</p> <div style="border: 1px solid black; height: 80px; margin: 10px 0;"></div>
3	$\frac{7}{12} + \left( \frac{3}{12} \right) =$	Multiplied fractions in the parentheses $\left( \frac{1}{4} \times \frac{3}{3} \right)$
4	$\frac{10}{12}$	Added 7 + 3 to get 10.

Figure 4. Example of a partially-worked problem with explanation.

Explanation tasks constitute the fourth type of assessment developed for the project. With respect to learning general concepts and principles, studies have shown self-explanation has significant effects on student learning (Chi, Bassok, Lewis, Reimman, & Glaser, 1989; VanLehn, 1996). In explanation tasks, students are expected to generate a clear, coherent explanation and how it can be used to solve problems and support their explanations with examples and illustrations, as in the example below:

In the space below, explain the distributive property and give several examples of how it can be used to solve problems or transform expressions.

A fifth type of problem is word problem. In these problems students generate a written solution or product in response to a problem solving prompt. The problem situation is described in text followed by a short (1–2 sentence) question or prompt, as shown below:

Sam and Jack deliver newspapers during the summer. One summer Sam made \$60 and saved \$20. The same summer, Jack made \$40 and saved \$15.

Who saved the greatest part of their summer earnings? Use fractions to explain how you solved the problem.

Another type of problem involves the use of graphics to introduce a problem situation. In these problems, students must use information presented in a diagram or picture to solve a problem (see example below).

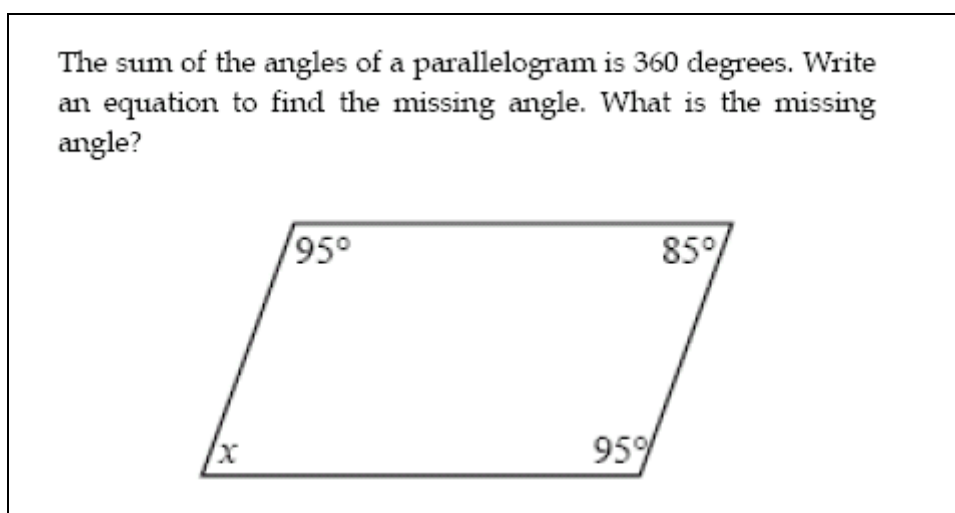


Figure 5. Example problem where a student must use information presented in the form of a diagram or picture to solve a problem.

A final problem type, in which tasks are embedded in an illustrated, comic book style narrative, was also developed. As well as investigating the use of research-based formative assessments in middle school classrooms, we decided to develop some alternative forms of assessments to test alongside the *Checks for Understanding*. The specific purpose in testing this type of assessment was to determine whether, if students find a math assessment more engaging, or more in keeping with the types of material they enjoy reading, they might perform better on the assessment. We are in the process of collecting initial pilot data on these narrative based assessments and will summarize those data elsewhere.

Table 1 shows the number of tasks piloted in each of these categories for the three primary domains we have been working in. About 200 additional tasks have been created and

not yet piloted. These additional tasks cover every domain and item type, and both sixth- and seventh-grade level content.

Table 1

Numbers of Tasks in Each Domain and Item Type Used in Pilot Testing

Item type	Solving equations	Distributive property	Rational number equivalence
Basic computational task	11	18	14
Partially-worked problems	5	8	4
Partially-worked problems with explanations	4	3	3
Word problems	10	0	4
Graphic problems	1	1	0
Explanation tasks	9	6	11
Tasks embedded into a narrative	6	10	0
<b>Total</b>	<b>46</b>	<b>46</b>	<b>36</b>

**3) Creating Forms for Pilot Testing:** Items developed as described above were grouped together into forms, called *Checks for Understanding*, designed to take around 15 minutes to complete. The amount of time devoted to POWERSOURCE<sup>®</sup> assessments was constrained by the districts we were working with. The districts determined that any assessment longer than around 15 minutes would be seen by teachers as a test, and would evoke complaints about too much district testing. However as it has turned out, this time frame actually has a number of advantages in focusing teachers and students' attention on students' understanding of a single concept and encouraging deep assessment without being too intrusive into or engendering teacher hostility about intrusion into instructional time. Longer assessments, for example district benchmark tests that occur every 9 weeks or so, tend to cover more content than teachers can deal with effectively after the test (e. g., if many students are deficient in many areas, there is not enough time to remediate all of them); shorter assessments given more often can be used more effectively.

Each *Check for Understanding* consisted of 3–5 items, with several different item types, with different cognitive demands, represented. For pilot testing, 112 items were compiled into 52 Checks or forms (14 forms for solving equations, 19 forms for properties of arithmetic, 17 forms for rational number equivalence, and 2 forms for the review and applications domain). We used an overlapping design across the forms (forms A and B had at

least two items in common, forms B and C had two common items, etc.) which ultimately allowed us to use IRT analyses to compare all items, and if necessary calibrate different forms in the future. In all cases the first two items on the test forms were basic computation/representation items. Subsequent items were either partially-worked problems with or without explanation, open-ended explanation tasks, or word or graphic problems (narrative tasks were tested separately.) Forms containing explanation tasks did not contain any other tasks besides the basic computational items. Most of the common items across forms were computation/representation tasks.

**4) Pilot testing the items:** In addition to determining whether students could read and understand the tasks, and whether the time estimates were reasonable (this was verified by teachers during follow-up discussions) the pilot tests were designed to compare the different item types with respect to the quality of information they provide, their diagnostic utility, and the information value they provide in relation to other tasks. For example, we were interested in the question: “Would some tasks provide more useful information about students’ knowledge and skill than others?” Finally, we expected that the pilot tests would provide useful information on student performance that could be used in the teacher resource materials and professional development.

### **Pilot testing procedures**

Pilot testing was carried out across 2 years (1 and 2). Each teacher participating in a pilot test received at least two different test forms. Two or more different forms were randomly assigned to the students within a classroom, and each teacher administered the assessments to two of their classrooms (if possible). Table 2 shows the number of teachers, classrooms and districts who participated in the pilot testing. Altogether we collected 3081 student responses across five school districts. Several teachers participated in multiple rounds of pilot testing.

Table 2.

Pilot-testing Numbers for Years 1 and 2 Combined

Years	Student responses	Teachers	Districts	Schools	Number of Test Forms
Pilot Years 1 and 2	3081	40	5	14	40

## Scoring Pilot Tests

Short-answer (one word, or one number answer) computation/representation tasks were scored dichotomously. Three- or four-point scoring rubrics were developed for all extended-response items. Students scored a 3 if their response indicated a mastery level understanding of the concept being evaluated, a 2 if their response indicated reasonable understanding of the concept, a 1 if their response indicated partial or incomplete understanding of the concept and a 0 if their response reflected an incorrect or no understanding of the concept. Descriptors for each score point are shown in Figure 5. Sample responses were also developed for the different score points.

Score point	Description
3	Student explains their answer in a way that demonstrates understanding of the big idea or key concept.
2	Student states a correct concept or procedure, but does not justify their answer or indicate understanding.
1	Student uses and/or states an irrelevant procedure or concept.
0	Student gives either no explanation or an incorrect explanation.

Figure 6. General scoring rubric

A total of eight expert teachers and three staff who were experienced in mathematics education and assessment scored the extended-response items, following procedures that The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) has validated in many previous studies (e.g., Baker, Aschbacher, Niemi, & Sato, 1992). The scoring session started with an orientation to the rubric and its application and practice with its application. This initial training period also provided the opportunity to refine the scoring rubrics and further customize them to the requirements of each individual item and/or for the constituent parts of each item. After initial training and refinement, rater reliability was established prior to starting actual rating. Raters all scored the same sample of 50 randomly selected test papers, to analyze initial concurrence and reliability. The results of the scoring of these papers were used to determine, using generalizability analyses, how many raters were needed to score each of the items as part of the full scoring session. On average, each item was scored by three raters, and all items were scored by at least two raters. The exact score-point agreement across all raters ranged from 84.5 to 100%, depending on the item. A sample task-specific scoring rubric is presented in Figure 7.

Explain why the following three fractions are equivalent.		
$\frac{5}{5}$	$\frac{121}{121}$	$\frac{1}{\frac{1}{2}}$

Score point	Description
3	The student response indicates: <ul style="list-style-type: none"> <li>∞ The fractions are equivalent because they are all equal to 1</li> <li>∞ Any number over itself is equal to 1</li> </ul>
2	The student response indicates that the fractions are the same because they have the same denominator and numerator. There is no mention that they are all equal to 1.
1	The student response indicates that the numbers can all divide into each other, or that they are all even numbers, or that they can all be reduced to the same fraction.
0	The student gives either no explanation or an incorrect/unintelligible explanation.

Figure 7. Example scoring rubric.

## Item Analyses

Several kinds of item-level analyses were carried out on the pilot test data. These include: confirmatory factor analyses, reliability analyses, and Item Response Theory (IRT) analyses. Our typical scheme for each set of *Checks for Understanding* form data was to first calculate reliability coefficients (Cronbach's alpha) for the selected items per each domain. Secondly, as another check of item quality, we conducted a principal component analysis and a confirmatory factor analysis for each test form to check whether the items exhibited the factor structure we expected (e. g., whether the computation items loaded on the same factor, etc). Thirdly, IRT analyses based on Rasch models were conducted in order to obtain item parameters (difficulties) and item characteristic and information curves. This information was then used to select items for future testing. The model-data fit was investigated using two model fit indices. One is the  $G^2$  index which is the Chi-square ( $\chi^2$ ) statistic and provided in PARSCALE phase 2 outputs, and the other is the mean square fit (MNSQ) statistics.

**Reliability and factor analysis:** The reliability of a test reflects the degree to which scores are free from random errors of measurement. Test reliability indicates the extent to which differences in test scores reflect real differences in the ability being measured and,



thus, the consistency of test scores across some change of condition, such as a change of test items or a change of time. For example, a low test-retest reliability coefficient means that a person's scores are likely to shift unpredictably from one time to another. The current reliability analysis estimates the internal consistency of a test form which is based on the inter-correlations among the items comprising the form. Results of reliability analyses allowed us to determine how much consistent information we were getting from particular items in a form and in turn make decisions on which items to include on the final *Checks for Understanding*. Test form names are abbreviated: Rational number equivalence (RNE), properties of arithmetic (PA), and solving equations (SE).

The reliability coefficients of the test forms used in the pilot study are provided in Table 3 and Appendix B. For instance, there were 13 PA forms tested each with 3–6 items. Given the relatively small number of items on each form, our reliability coefficients were lower than they would be if each form contained more items. Nevertheless, the reliability estimates for the 13 PA forms ranged from 0.371 to 0.794. Two test forms (RNE-14, and SE-3) had negative values of alpha, which seemed to be caused by negative average covariance among items. The results of principal component analysis indicated that a single factor could explain at least 29.58% of the total variance. In case of Year 2 SE forms, the extended items were not scored so the reliability estimates reported in Appendix B often appeared to be very small or negative. Principal component analysis was not executed for these forms.

Table 3

Reliability Coefficients and Goodness-of-fit of a Single Factor Model for *Checks for Understanding* Forms in Three Domains (Pilot Study).

Year	Test form	Reliability	% of Variance*
Year 1	RNE-5 ( $N = 55$ )	0.687	35.826
	RNE-7-5 ( $N = 142$ )	0.535	30.418
	RNE-8 ( $N = 77$ )	0.597	45.572
	RNE-A4 ( $N = 127$ )	0.674	50.711
	RNE-B4 ( $N = 129$ )	0.707	53.511
	RNE-C3 ( $N = 124$ )	0.718	47.371
Year 2	RNE-13 ( $N = 107$ )	0.613	56.822
	RNE-14 ( $N = 98$ )	-0.094	52.255
	RNE-15 ( $N = 101$ )	0.788	50.766
	RNE-16 ( $N = 74$ )	0.711	47.123
	RNE-17 ( $N = 88$ )	0.741	50.148
Year 1	PA-1 ( $N = 58$ )	0.443	43.272
	PA-2 ( $N = 114$ )	0.689	45.437
	PA-3 ( $N = 52$ )	0.558	43.233
	PA-6 ( $N = 25$ )	0.558	49.657
	PA-7 ( $N = 55$ )	0.551	42.781
	PA-8 ( $N = 57$ )	0.371	29.658
	PA-9 ( $N = 54$ )	0.682	46.800
	PA-10 ( $N = 135$ )	0.725	54.875
	PA-11 ( $N = 131$ )	0.583	39.967
	PA-12 ( $N = 43$ )	0.785	54.080
	PA-13 ( $N = 84$ )	0.526	51.755
	PA-14 ( $N = 82$ )	0.794	55.847
	PA-15 ( $N = 28$ )	0.643	41.980
Year 2	SE-1 ( $N = 87$ )	0.470	34.627
	SE-2 ( $N = 80$ )	0.561	39.041
	SE-3 ( $N = 17$ )	-0.226	40.703
	SE-4 ( $N = 59$ )	0.545	37.645
	SE-5 ( $N = 78$ )	0.512	35.838
	SE-7 ( $N = 121$ )	0.712	40.531
	SE-8 ( $N = 109$ )	0.687	41.402

*Note.* \*Percentage of variance explained by the main principal component. RNE = rational number equivalence, PA = properties of arithmetic, SE = solving equations.

Several items had extremely low  $p$  values (difficulty index of classical test theory), around 0.03, and when these items were removed from the analysis, there was a significant increase in Cronbach's alpha for that particular form. We observed this pattern in many of the partially-worked problems with and without explanations. The effect was strongest for the worked examples requiring explanations. These were the items where students had to describe the next step of a problem and include an explanation and justification for that next step. In our discussion with teachers we learned that they did not use worked examples in their problem-solving instruction. Based on these findings we decided to omit this type of partially-worked examples from the final forms of the *Checks for Understanding* and to incorporate them instead into the instructional resources.

Detailed analyses of each of the *Checks for Understanding* forms were conducted. By way of an example, the detailed results of the reliability analysis and component loadings for two of the PA forms are presented in Table 4. See Appendix C for the items included on these two forms.

Table 4  
Reliability and Factor Analysis for Two *Checks for Understanding* Forms Including in Pilot Testing.

Test form			Reliability
Items	Short/Extended	Component loading	Alpha if item deleted
<b>PA-1 (<math>n = 58</math>)</b>			<b>0.443</b>
PA-BT-3	short item	0.688	0.260
PA-BT-4	short item	0.711	0.320
PA-PW-3a	short item	0.761	0.268
PA-PW-3b	short item	0.639	0.224
PA-PWE-1	extended item	-0.446	0.682
<b>PA-2 (<math>n = 114</math>)</b>			<b>0.689</b>
PA-BT-3	short item	0.485	0.675
PA-BT-4	short item	0.403	0.693
PA-PW-2a	short item	0.828	0.576
PA-PW-2b	short item	0.909	0.550
PA-PW-2c	short item	0.902	0.568
PA-PWE-2	extended item	0.041	0.765

Note. PA = properties of arithmetic.

**Test Form PA-1:** PA-1 consisted of four assessment items, two basic tasks (PA-BT-3 and P-BT-4), one partially-worked problem with two parts (PA-PW-3a and PA-PW-3b) and one partially-worked problem with an explanation (PA-PWE-1). The reliability coefficient (Cronbach's alpha) for the form was 0.443 and the items seem to hang together well and removing any of the items (except PA-PWE-1) from the form will impair the test reliability. PA-PWE-1 also appeared to have negative loading value (-0.446), which meant this item measured something different from the main construct. The results of the factor analysis indicate that there is one underlying factor assessed by this test form (except PA-PWE-1).

**Test Form PA-2:** Form PA-2 consisted of four items, two basic tasks (PA-BT-3 and PA-BT-4), one partially-worked problem with three parts (PA-PW-2a, b, & c) and one partially-worked problem with an explanation (PA-PWE-2). The reliability coefficient for the form was 0.689. The results of the principal component analysis show that the main component can be highly related to each item except PA-PWE-2. Removal of this item would increase the reliability from 0.689 to 0.765, which indicated that the item PA-PWE-2 impaired the internal consistency of this test form.

**Item Response Theory (IRT) Analyses.** All items were analyzed using Rasch models to quantify differences in item or category difficulties. On a probit difficulty scale (i.e.,  $D = 1.7$  was used in Rasch models), the difficulties of all the tasks were determined. The item parameter calibration runs were conducted using PARSCALE (Muraki & Bock, 1997) under the one parameter logistic model for dichotomous items and partial credit model (PCM; Masters, 1982) for extended response items. Because the former model is a special case of the latter, the Rasch model used in this report can be expressed as follows. The probability that an examinee  $j$  scores  $z$  with  $z = 0, 1, \dots, Z_i$  on item  $i$  with  $Z_i + 1$  response categories is

$$P(z | \theta_j, \beta_i, \tau_{ci}) = \frac{\exp \sum_{c=0}^z D[\theta_j - (\beta_i - \tau_{ci})]}{\sum_{y=0}^{Z_i} \exp \sum_{c=0}^y D[\theta_j - (\beta_i - \tau_{ci})]}, \quad (1)$$

where  $\beta_i$  denotes the difficulty of item  $i$ , and  $\tau_{ci}$  represents the location parameter for a category on item  $i$ . For model identification, Equation (1) needs to set  $\tau_{0i} = 0$ ,  $\sum_{c=1}^{Z_i} \tau_{ci} = 0$  and

$$\exp \sum_{c=0}^0 \alpha_i [\theta_j - (\beta_i - \tau_{ci})] = 1.$$

To check the model-data fit, as aforementioned, two statistics were used. One is the  $G^2$  provided in PARSCALE with which a significance test can be executed to decide if the model statistically fits a given item data based on  $\chi^2$  distribution. The other is MNSQ which is used to determine whether items were functioning in a way that is congruent with the assumptions of the Rasch model. Two types of MNSQ values are presented, OUTFIT and INFIT. MNSQ OUTFIT values are sensitive to outlying observations. MNSQ INFIT values are sensitive to behaviors that affect students' performance on items near their ability estimates. They are the chi-square statistics divided by its degrees of freedom. Consequently its expected value is close to 1.0. Values greater than 1.0 (underfit) indicate unmodeled noise or other source of variance in the data—these degrade measurement. Values less than 1.0 (overfit) indicate that the model predicts the data too well. According to the item analysis specification, the model is considered to be moderately misfit if the values are between 1.5 and 2.0 and highly misfit if the values are greater than 2.0.

The results of the model-data fit analyses for the pilot study are provided in Appendices D-1 (RNE), D-3 (PA) and D-5 (SE). For example, as shown in Appendix D-3, 15 items among the total 37 items in PA (pilot a) study had  $p$ -values less than 0.05 in the  $G^2$  statistic analysis. The other items appeared to fit the Rasch models. Because it is known that the  $G^2$  can control the type I error rates only in very limited testing conditions (Orlando & Thissen, 2000), however, the fit of Rasch models seemed to be interpreted using the concept of INFIT and OUTFIT more appropriately. According to the INFIT and OUTFIT values, 15 and 16 items appeared to have MNSQs larger than 1.0, respectively. It was rare to find highly misfit items, however, because most items in PA pilot test forms had MNSQ, INFIT, and OUTFIT values less than 1.5.

The item parameter calibration was executed for each domain rather than for a form in each domain, whereby the item parameter estimates of a domain in pilot testing could be considered being put onto a common scale. Error estimates are given in the parenthesis following each difficulty estimate. For example, as shown in Appendix D-1, the items in the RNE domain had item difficulties ranging from -3.10 (0.54) to 1.61 (0.15). Also, the item difficulty parameter estimates for PA in (pilot A) ranged between -1.60 (0.18) and 2.06 (0.19) as provided in Appendix D-3.

Where test items contained multiple parts, these parts were split apart for the IRT analysis. For instance, for a partially-worked problem, we separated out each component of a problem and analyzed it separately (see Figure 8).

	Problem Solving Step
1	$x + 5 = 10$
2	$x + 5 - 5 = 10 - \square$
3	$x = \square$

Figure 8. Partially-worked problem with two components.

This allowed us to determine the difficulty of each component of a multi-part problem and determine if certain steps were more or less difficult for students. The IRT analyses allowed us to quantify the differences in item difficulty across all items, and enabled us to calibrate different assessment forms. Using this information we were able determine how much information the selected items provide about student knowledge in a particular domain and the extent to which different levels of student ability interact with each of the selected items.

Based on the results from the IRT analyses, we drew item characteristic and information curves for each item characteristic and information curves for a set of items. For each set of items within a domain, the item characteristic curve gave information on how well students of different ability levels performed on different items. In Appendix E, the item (category) characteristic curves of items in each domain's field test forms are provided. The item information curve also shows how much item contribution at each ability level is expected for accurate ability estimation. The IRT results described above gave us information on the difficulty or category parameters of each item, which was used in conjunction with frequency data and factor analysis as additional criteria to determine which items to use in field testing.

### Item Analyses for Field Testing

As explained earlier in this report, based on the multiple criteria discussed above we used pilot-test information to refine our assessment forms and develop *Checks for Understanding* that then were then administered to students as part of a field test of the instructional sensitivity of the *Checks* as well as the effectiveness of the professional development and instructional materials in each of the four conceptual domains.

Our POWERSOURCE<sup>®</sup> field test had two major objectives: a) to refine the design of our materials and b) to test POWERSOURCE effectiveness. We employed a randomized, controlled design to address the following specific research questions: Does the use of POWERSOURCE<sup>®</sup> formative assessments improve student performance on assessment of the key mathematical ideas and on relevant subscales of the state assessment, relative to the performance of a comparison group? Data from the field test also gave us useful information on the *Checks for Understanding* and the items within them. A description of the field test design follows (more details of the results of this study can be found in Choi, Phelan, Niemi & Vendlinski, in progress).

### Sites and Design

We field tested both the POWERSOURCE<sup>®</sup> assessments and associated instructional materials at several different sites. Fifty-eight teachers were recruited from 25 middle schools in Arizona (two districts: AZ-1 and AZ-2) and California (two districts: CA 1 and CA 2). Table 5 provides information on the study participants and districts.

Table 5  
Participants in the POWERSOURCE<sup>®</sup> Field Test

	Students	Teachers	Districts	Number of test forms
Field Testing	2340	58	4	10

Within each district, teachers were randomly assigned to experimental POWERSOURCE<sup>®</sup> and comparison groups, but the definition of treatment varied in response to both local district needs and our intent to try out different ways of using our materials. Teachers in each district were randomly assigned to two groups. Experimental group teachers in all cases participated in initial summer professional development and after school follow-up sessions, and used project modules, including the *Checks for Understanding* and instructional supports, but comparison group experiences varied. District CA-1 represented the cleanest design for examining the effects of the POWERSOURCE<sup>®</sup> intervention in total. Here, the comparison group received no POWERSOURCE<sup>®</sup> professional development (although teachers did participate in usual district professional development for mathematics) and had no access to instructional supports, although teachers were asked to administer the *Checks for Understanding* for use as dependent variable. In the other three districts, the comparison group participated in POWERSOURCE<sup>®</sup> professional development and administered the *Checks for Understanding*, but had no access to the instructional supports, in effect providing

a test of the value added by the instructional support. All teachers gave eight *Checks for Understanding* throughout the school year, two for each module as described above.

### **Reliability and Factor Analysis**

For the field test, we added an additional domain to the existing three POWERSOURCE<sup>®</sup> domains (RNE, SE and PA). This domain (review and applications: RA) included review items and items testing the application of the core principles in the other three domains. Table 6 contains the reliability coefficients for *Checks for Understanding* forms used in the field test. Each domain had two test forms—for a total of 8 forms. The range of reliability coefficients (Cronbach's alpha) was between 0.554 and 0.863. Table 3 also includes the results of a confirmatory single factor analysis such as root mean square error of approximation (RMSEA), goodness of fit index (GFI). Also, the percentages of variance explained by the main principal component are provided. Based on the inter-correlations between the items, factor analysis further determines the theoretical constructs that might be represented by the set of items in a form. This analysis allowed us to look at the *Checks for Understanding* forms in each domain and see if the items exhibited the factor structure we expected. Values less than .05 for the RMSEA indicate a close fit, with values as high as .08 representing a reasonable fit (Joreskog & Sorbom, 1993, p. 124). The goodness of fit (GFI) provides a measure of the relative amount of variance and covariance accounted for by the model. Values greater than .90 for the GFI measure are required to indicate a good fit (Byrne, 1994). According to these criteria, the RMSEA and GFI values in Table 6 indicate that the items' variance in each test could be explained well by a single construct. In each form, the main component accounted for 27.529% through 50.195% of the total variance, that suggested the items measured a uni-dimensional trait.



Table 6

Reliability Coefficients and Goodness-of-fit of a Single Factor Model for *Checks for Understanding* Forms in Four Domains (Field Test)

Test form	Reliability	RMSEA	GFI	% of Variance*
RNE-10 ( $n = 3,320$ )	0.668	0.021	0.998	35.682
RNE-9 ( $n = 3,068$ )	0.554	0.024	0.996	27.529
PA-18 v2 ( $n = 3,101$ )	0.812	0.019	0.996	43.896
PA-19 v2 ( $n = 3,068$ )	0.827	0.021	0.988	44.328
SE-11 v2 ( $n = 2,978$ )	0.641	0.051	0.996	37.470
SE-12 v1 ( $n = 2,961$ )	0.718	0.059	0.989	39.349
RA-1 v3 ( $n = 1,163$ )	0.863	0.057	0.958	40.081
RA-2 v2 ( $n = 1,111$ )	0.831	0.057	0.985	50.195

*Note.* \*Percentage of variance explained by the main principal component. RNE = rational number equivalence, PA = properties of arithmetic, SE = solving equations, RA = review and applications.

The more detailed results of the reliability analysis and component loadings for two field test PA forms are presented in Table 7. See Appendix E for the items included on these two forms.

Table 7

Reliability and Factor Analysis for Two *Checks for Understanding* Forms Including in Field Testing.

Test form			Reliability
Items	Short/Extended	Component loading	Alpha if item deleted
<b>PA-18 (<math>n = 3,101</math>)</b>			<b>0.812</b>
PA-BT-1	short item	.637	0.792
PA-BT-13	short item	.695	0.780
PA-BT-9	short item	.649	0.790
PA-PW-8a	short item	.691	0.787
PA-PW-8b	short item	.781	0.772
PA-PW-8c	short item	.766	0.775
PA-EX-1a	extended item	.260	0.837
PA-EX-1b	extended item	.678	0.788
<b>PA-19 (<math>n = 3,068</math>)</b>			<b>0.827</b>
PA-BT-2	short item	.726	0.800
PA-BT-3	short item	.717	0.800
PA-BT-12	short item	.695	0.802
PA-PW-4-a	short item	.759	0.797
PA-PW-4-b	short item	.869	0.783
PA-PW-4-c	short item	.837	0.786
PA-PWE-3	extended item	.362	0.837
PA-EX-6a	extended item	.476	0.826
PA-EX-6b	extended item	.276	0.844

Note. PA = properties of arithmetic.

### **Check for Understanding PA-18:**

*Check for Understanding* form PA-18 consisted of five items: three basic tasks (PA-BT-1, PA-BT-13, & PA-BT-9), one partially-worked problem with three parts (PA-PW-8-a, b, & c), and one explanation task with two parts (PA-EX-1-a, & b). The reliability coefficient (Cronbach's alpha) for the form was 0.812. The items hung together well and removing any of the items (except PA-EX-1a) from the form impaired the test reliability. PA-EX-1a also appeared to have relatively small loading value (.260). PA-EX-1a is an item that asked students to explain the distributive property. Other analyses have shown us that explanation items are difficult for students. Indeed, results from the IRT analysis indicate a difficulty estimate of 1.61 (0.03) for this item, but 0.34 (0.02) for the second part of this question (PA-EX-1b) that asks students to give an example of how to use the property. According to the

factor analysis the explanation task is measuring a different construct than the rest of the items and that aside from this item there is one underlying factor assessed by Form PA-18.

#### ***Checks for Understanding PA-19:***

*Check for Understanding* Form PA-19 consisted of six items: three basic tasks (PA-BT-2, PA-BT-3, & PA-BT-12), one partially-worked problem with three parts (PA-PW-4-a, b, & c), one partially-worked problem with an explanation (PA-PWE-3), and one explanation task with two parts (PA-EX-6-a, & b). The reliability coefficient for the form was 0.827. The results of principal component analysis show that the main component can be highly related to each item except for two items. PA-EX-6b had a small loading value (0.276). Again, this item is one requiring students to provide an explanation of the distributive property. PA-PWE-3 also had a small loading value (0.362) and was task requiring students to provide an explanation. Implications of these findings will be discussed further. Considering the confirmatory factor analysis results together, there seems to be a single construct measured by this form.

Principal component analyses were carried out on the other three domains (RNE, SE, & RA). See Appendix F for detailed results of the reliability analysis and component loadings.

**Item Response Theory (IRT) Analyses.** Appendices D-2 (RNE), D-4 (PA), D-6 (SE), and D-7 (RA) contain item parameters of dichotomous and polytomous items estimated using the program PARSCALE. Also, the item characteristic curve of every item in the five domains is provided in Appendix E. The item parameter calibration for field testing was executed for each domain rather than for a form in each domain, whereby the item parameter estimates of a domain could be considered being put onto a common scale. For example, item parameter estimates for SE in field testing ranged from -0.21 (0.03) to 1.58 (0.04), see Appendix D-6. Error estimates are given in the parenthesis following each difficulty estimate. Also, item parameters of pilot tests years A and B and the field test were estimated separately. Because there exist some common items between them within each domain, however, it will be possible to link the item parameters using test characteristic method (Stocking & Lord, 1983) or mean/sigma method (Marco, 1977) if required.

Figure 9 shows the item (category) characteristic curves and item information curves of items in the SE-11 field test form. SE-11 has five dichotomous and only one polytomous items. Figures 9a, 9b, and 9c show item characteristic curves of the dichotomous items (SE-BT-7, SE-BT-8, SE-PW-3a, SE-PW-3b, & SE-PW-3c), item category characteristic curves of the polytomous item (SE-EX-9), and item information curves of the six items, respectively.

We can interpret these item characteristic and information curves under Rasch models to indicate that two of the assessment items SE-BT-7 and SE-BT-8 (shown in Figure 10) are less difficult items and as such provide more information for low ability students than high ability students. Conversely, the item SE-PW-3b (shown in Figure 11) shows the opposite pattern. This item provides more information about the high ability students than about the lower ability students as shown in Figure 9c. In Figure 9b, it can be inferred that the students with ability less than about  $\text{Theta} = 0.8$  had a high probability of scoring zero on this item. As shown in Figure 9c, the polytomous item (SE-EX-9; shown in Figure 12) was able to grant much more information than any single dichotomous item, and provided more information for the students with higher ability.

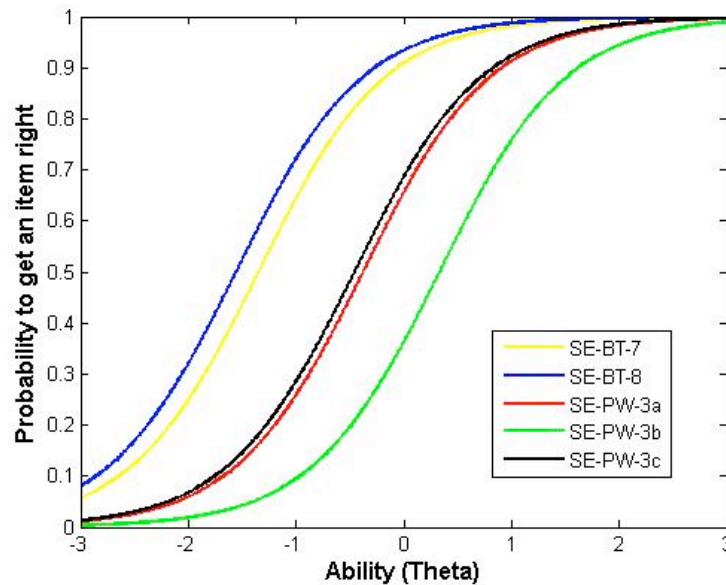


Figure 9a. Item (Category) Characteristic Curves and Item Information Curves for SE field test items.

Item Characteristic Curves of the five dichotomous items.

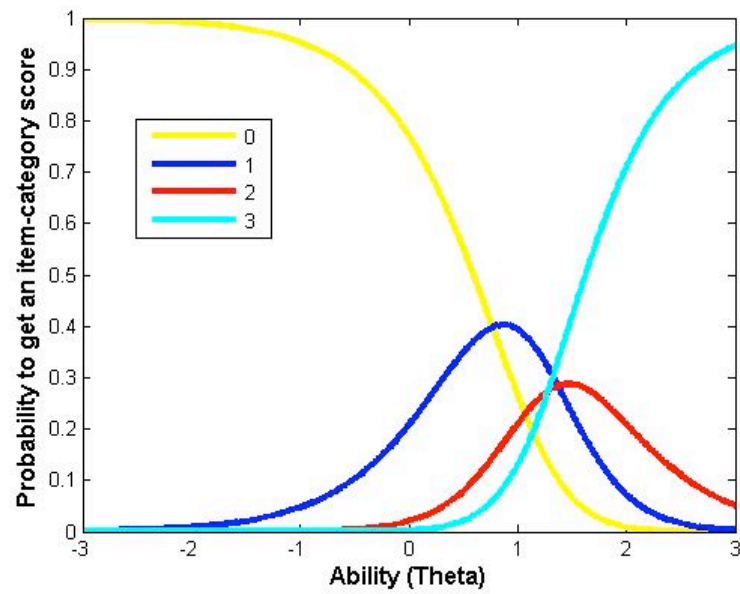


Figure 9b. Item (Category) Characteristic Curves and Item Information Curves for SE field test items.

Item Category Characteristic Curves of the polytomous item (SE-EX-9).

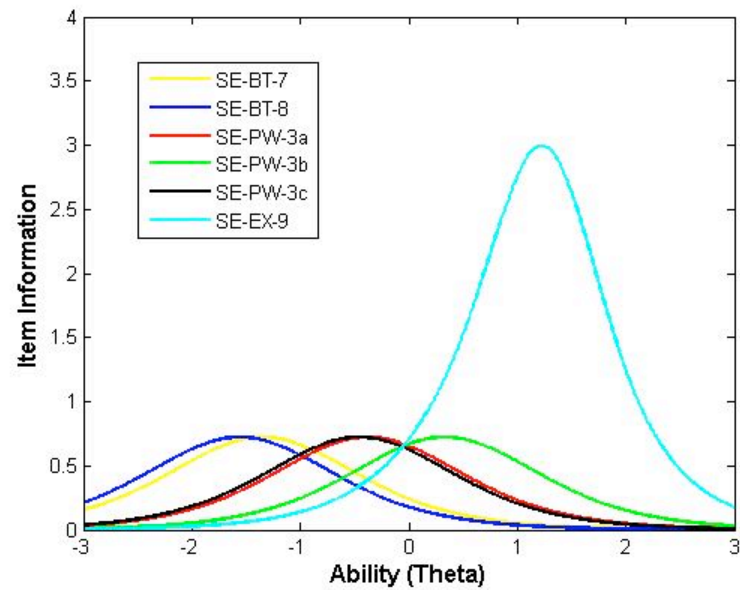


Figure 9c. Item (Category) Characteristic Curves and Item Information Curves for SE field test items.

Item Information Functions of the six items.

**1**     $x - 5 = 15$

**2**     $12 = x + 4$

Figure 10. Basic computational tasks included on Check for Understanding Form SE-11.

**3** A student was asked to solve for the value of  $x$  in the diagram below.

This is how the student set up the problem to find the missing angle ( $x$ ). Can you fill in the missing numbers in step 2, 3, and 4?

	Problem Solving Step	
<b>1</b>	$90 + 60 + x = 180$	
<b>2</b>	<div style="border: 1px solid gray; display: inline-block; width: 30px; height: 20px;"></div> $+ x = 180$	SE-PW-3 a
<b>3</b>	$150 - 150 + x = 180 - $ <div style="border: 1px solid gray; display: inline-block; width: 30px; height: 20px;"></div>	SE-PW-3 b
<b>4</b>	$x = $ <div style="border: 1px solid gray; display: inline-block; width: 30px; height: 20px;"></div>	SE-PW-3 c

Figure 11. Partially-worked problem included on Check for Understanding Form SE-11.

**4** Your teacher asks you to explain to another student how to solve an equation like this one:  $x - 5 = 43$ . In your answer, be sure to explain the goal of solving this equation and the principles you would use to solve the equation.

$x - 5 = 43$

Figure 12. Explanation item included on Check for Understanding SE-11.

The results of the model-data fit analyses for the field testing are provided in Appendix D. For example, as shown in Appendix D-2, only three items had  $p$ -values larger than 0.05 in the  $G^2$  statistic analysis among the 14 items on the RNE field-test forms.

The other items appeared to be misfitting the Rasch models. Because it is known that the  $G^2$  can control the type I error rates only in very limited testing conditions (Orlando & Thissen, 2000), however, the fit of Rasch models seemed to be interpreted using the concept of INFIT and OUTFIT more appropriately. In the field test, every item in RNE forms had MNSQ INFIT value less than 1.5 and only three items (RN-BT-7, EX-9, & EX-6b1) showed some misfit problems in terms of MNSQ OUTFIT. MNSQ INFIT values provide model item fit information around the difficulty parameter where the item is most informative, the appropriate fit is very important in this area. The items in RNE field testing appeared to have INFIT values less than 1.24. A similar pattern appeared in the other domains' field test forms.

Even though a few items appeared not to be fit by Rasch models, this does not invalidate the measure. This simply indicates that beyond the strong overall achievement measured by each domain's test forms, there are also some minor dimensions of achievement that impact the individual item scores of individual students. That the overall dimensions (or principal components) measured by each subject assessment are very strong is demonstrated by both (a) strong Cronbach's alpha internal consistency reliabilities (a measure of measurement precision of the overall dimension derived outside the IRT model), and (b) the positive results from the confirmatory factor analysis and principal component analysis.

## **Discussion**

Results of the pilot testing reported here suggest that relatively brief formative assessments focused on key conceptual domains can provide reliable and useful information on students' levels of understanding, as measured by explanation tasks, and related skills, including problem solving, computation, and use of symbolic representations. Item design began with an analysis of middle school mathematics in terms of major organizing concepts and skills related to those concepts; this analysis in turn drew both on a set of big ideas elicited from mathematicians and on content delineated in state standards. We then developed 15-minute multi-item *Checks for Understanding* that also showed reasonable reliability and information value when administered in instructional settings. Of particular note are the results for some of the explanation tasks administered on our *Check for Understanding* forms. In most cases, these items provided us much more information than some of the other dichotomous items and were better able to distinguish those students with

high ability from those with lower ability levels. We know from both the literature and discussion with teachers that tasks requiring students to explain their thinking are difficult for students (Lester, 1994). A NAEP report focused on problem solving indicated that the average percentage of students giving satisfactory responses (or better) was around 16% at Grade 4 and only 8% at Grade 8 (Dossey, Mullis, & Jones, 1993). Indeed, many teachers in this study reported serious deficiencies in their students' abilities to problem solve in this way, partly as a result of their lack of opportunity to do so. State and other assessments tend to have very few (if any) explanation tasks.

The results on the explanation tasks confirm that these types of tasks are indeed difficult, particularly for the lower achieving students, but also reaffirm that we can get a great deal of information about student understanding using this type of task. Used in concert with other—more basic, computational tasks, we can get a better overall picture of student understanding—in particular of the deeper concepts we are hoping students will gain as part of the POWERSOURCE<sup>®</sup> project. Many students can execute algorithms and carry out basic calculations, but then have difficulties explaining the concepts underlying them. Thereby illustrated a shaky understanding, at best, of the material covered. This information is critical as we further develop and test our instructional resources for teachers. Clearly, possessing the knowledge to solve a simple equation, does not necessarily equip one to explain why an equation is solved in a particular way. We do not expect these skills will be developed without explicit instruction, opportunity to practice and exposure. Thus we will continue to incorporate instruction on these tasks into the instructional resources provided to teachers and into our *Checks for Understanding*.

These results, however, are just part of the evidence needed to validate the tasks as formative assessments. Other evidence includes information on the sensitivity of the tasks to instruction (so that they are not just measuring, for example, general intelligence or mathematics achievement) and the utility of the tasks in a formative assessment system, which means that teachers are able to use the assessments to make more informed and effective instructional decisions. We have obtained significantly positive evidence on the instructional sensitivity of the tasks in experimental studies (e.g., Choi et al., in progress), and we are currently conducting large-scale experimental (i.e., with random assignment to treatments) studies of the value of the *Checks for Understanding* as formative assessments in the POWERSOURCE<sup>®</sup> program.



## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice*, 36(4), 247–254.
- Baker, E. L., Aschbacher, P. R., Niemi, D. & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations*. (CRESST Tech. Rep. No. 652). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Englewood Cliffs, NJ: Prentice-Hall.
- Black, P. J., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5(1), 7–74.
- Black, P. J., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: School of Education, King's College. (See also article with the same title, 1998, in *Phi Delta Kappan*, 80(2), 139–148.)
- Black, P. J., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, UK: Open University Press.
- Brown, R. S. & Niemi, D. N. (2007). Investigating alignment of high school and community college assessments in California. San Jose, CA: National Center for Public Policy in Higher Education.
- Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Thousand Oaks, CA: Sage Publications.
- California Department of Education. (2007). *Standardized testing and reporting (STAR) program*. Retrieved October 16, 2008, from <http://star.cde.ca.gov/star2007/viewreport.asp?ps=true&lstTestYear=2006&lstTestType=C&lstCounty=&lstDistrict=&lstSchool=&lstGroup=1&lstSubGroup=1>
- Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.
- Choi, K. C., Phelan, J., Niemi, D. N., & Vendlinski, T. (in progress). The effects of POWERSOURCE<sup>®</sup> on student performance. (Draft Deliverable to IES, Contract No. R305A050004). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Dossey, J. A., Mullis, I. V. S., & Jones, C. O. (1993). *Can students do mathematical problem solving? Results from constructed-response questions in NAEP's 1992 mathematics assessment*. Washington, DC: National Center for Education Statistics.

- Joreskog, K. A., & Sorbom, D. (1993). *LISREL 8: A guide to the program and applications*. Chicago: Scientific Software Inc.
- Lester, F. K. (1994). Musings about mathematical problem-solving research: 1970–1994. *Journal for Research in Mathematics Education*, 25, 660–675.
- Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14, 139–160.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Minstrell, J. (2000). Student thinking and related assessment: Creating a facet assessment-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), Committee on the Evaluation of National and State Assessments of Educational Progress, National Research Council. (2000). *Grading the nation's report card: Research from the evaluation of NAEP. Commission on Behavioral and Social Sciences and Education*, (pp. 44–73). Washington, DC: National Academy Press.
- Muraki, E. & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data*. Chicago: Scientific Software.
- Niemi, D. N. (1996). *Instructional influences on content area explanations and representational knowledge: Evidence for the construct validity of measures of principled understanding*. (CRESST Tech. Rep. No. 403). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Niemi, D. N., Baker, E. L., & Sylvester, R. (2007). Scaling up, scaling down: seven years of performance assessment development in the nation's second largest school district. *Educational Assessment*, 12, 195–214.
- Niemi, D. N., Vallone J. & Vendlinski, T. (2006). *The power of big ideas in mathematics education: Development and pilot testing of POWERSOURCE<sup>®</sup> assessments* (CRESST Tech. Rep. No. 697). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Niemi, D. N. & Phelan, J. (2000, April). *Creating a cognitive blueprint for assessment design*. Presentation at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Renkl, A., Atkinson, R. K., & Maier, U. H. (2000). From studying examples to solving problems: Fading worked-out solution steps helps learning. In L. Gleitman & A. K. Joshi (Eds.), *Proceeding of the 22nd Annual Conference of the Cognitive Science Society* (pp. 393–398). Mahwah, NJ: Lawrence Erlbaum Associates.

- Stiggins, R. J. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, 86(1), 22.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Sweller, J., van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10, 251–296.
- VanLehn, K. (1996). Cognitive skill acquisition. In J. Spence, J. Darly & D. J. Foss (Eds.), *Annual review of psychology* (Vol. 42, pp. 513–539). Palo Alto, CA: Annual Reviews.



## Appendix A

### A1. RNE: Year 1 Pilot Study

FORM	ITEM	Short/Extended	Etc.	Reliability	Alpha if item deleted
RNE-5 (N = 55)	rnebt4	short item			0.647
	rnebt6	short item			0.661
	rnpw2a	short item			0.696
	rnpw2b	short item		0.687	0.611
	rnpw2c	short item			0.612
	<i>rnepwe1_a</i>	<i>extended item</i>			0.664
	<i>rnepwe1_b</i>	<i>extended item</i>			0.670
RNE-7-5 (N = 142)	rnebt1	short item			0.518
	rnebt6	short item			0.504
	<i>rneex6_a</i>	<i>extended item</i>	<i>rneex6_a(i)</i>	0.535	0.480
	<i>rneex6_b</i>	<i>extended item</i>	<i>rneex6_a(ii)</i>		0.466
	<i>rneex6_c</i>	<i>extended item</i>	<i>rneex6_b(i)</i>		0.508
	<i>rneex6_d</i>	<i>extended item</i>	<i>rneex6_b(ii)</i>		0.460
RNE-8 (N = 77)	rnebt2	short item			0.453
	rnebt5	short item			0.567
	<i>rnewp2_a</i>	<i>extended item</i>		0.597	0.582
	<i>rnewp2_b</i>	<i>extended item</i>			0.493
RNE-A4 (N = 127)	rnebt1	short item			0.635
	rnebt2	short item			0.622
	<i>rneex1_a</i>	<i>extended item</i>		0.674	0.560
	<i>rneex1_b</i>	<i>extended item</i>			0.612
RNE-B4 (N = 129)	rnebt2	short item	not “rnebt3”		0.635
	rnebt4	short item			0.686
	<i>rneex2_a</i>	<i>extended item</i>		0.707	0.679
	<i>rneex2_b</i>	<i>extended item</i>			0.565
RNE-C3 (N = 124)	rnebt5	short item			0.669
	rnebt6	short item			0.717
	<i>rneex3_a</i>	<i>extended item</i>		0.718	0.665
	<i>rneex3_b</i>	<i>extended item</i>			0.627
	<i>rneex3_c</i>	<i>extended item</i>			0.669

## A2. RNE: Year 2 Pilot Study

FORM	ITEM	Short/Extended	Etc.	Reliability	Alpha if item deleted
RNE-13 ( <i>N</i> = 107)	RN-BT-10	short item			0.534
	RN-BT-4	short item		0.613	0.311
	RN-BT-8	short item			0.655
RNE-14 ( <i>N</i> = 98)	RN-BT-1	short item		-0.094	
	RN-BT-8	short item			
RNE-15 ( <i>N</i> = 101)	RN-BT-5	short item			0.798
	RN-BT-8	short item			0.768
	RN-PW-2a	short item		0.788	0.731
	RN-PW-2b	short item			0.710
	RN-PW-2c	short item			0.716
	RN-BT-7	short item	not “RN-EX-12a”		0.799
RNE-16 ( <i>N</i> = 74)	RN-BT-6	short item			0.692
	RN-BT-8	short item			0.711
	RN-PW-3a	short item		0.711	0.624
	RN-PW-3b	short item			0.661
	RN-PW-3c	short item			0.619
RNE-17 ( <i>N</i> = 88)	RN-BT-1	short item			0.742
	RN-BT-2	short item			0.678
	RN-PW-4a	short item		0.741	0.640
	RN-PW-4b	short item			0.746
	RN-PW-4c	short item			0.661

### A3. PA: Year 1 Pilot Study

FORM	ITEM	Short/Extended	Reliability	Alpha if item deleted
PA-A v3 (PA-1 v3) (N = 58)	PA-BT-3	short item	0.443	0.260
	PA-BT-4	short item		0.320
	PA-PW-3a	short item		0.268
	PA-PW-3b	short item		0.224
	PA-PWE-1	extended item		0.682
PA-2 v4 & PA-B v4 (N = 114)	PA-BT-3	short item	0.689	0.675
	PA-BT-4	short item		0.693
	PA-PW-2a	short item		0.576
	PA-PW-2b	short item		0.550
	PA-PW-2c	short item		0.568
	PA-PWE-2	extended item		0.765
PA-C v2 (PA-3 v2) (N = 52)	PA-BT-3	short item	0.558	0.540
	PA-BT-4	short item		0.448
	PA-EX-1a	extended item		0.521
	PA-EX-1b	extended item		0.426
PA-F v3 (PA-6 v6) (N = 25)	PA-BT-5	short item	0.558	0.537
	PA-BT-6	short item		0.692
	PA-PW-4a	short item		0.287
	PA-PW-4b	short item		0.264
	PA-PW-4c	short item		0.255
	PA-PWE-1	extended item		0.719
PA-G v3 (PA-7 v3) (N = 55)	PA-BT-3	short item	0.551	0.486
	PA-BT-4	short item		0.510
	PA-EX-2a	extended item		0.408
	PA-EX-2b	extended item		0.506
PA-H v3 (PA-8 v3) (N = 57 )	PA-BT-3	short item	0.371	0.398
	PA-BT-7	short item		0.136
	PA-PW-5a	short item		0.332
	PA-PW-5b	short item		0.370
	PA-PWE-3	extended item		0.331

(table continues)

### A3. PA: Year 1 Pilot Study (continued)

FORM	ITEM	Short/Extended	Reliability	Alpha if item deleted
PA-9 v3 (N = 54)	PA-BT-8	short item	0.682	0.642
	PA-BT-9	short item		0.734
	PA-PW-4a	short item		0.538
	PA-PW-4b	short item		0.538
	PA-PW-4c	short item		0.546
	<i>PA-PWE-3</i>	<i>extended item</i>		<i>0.760</i>
PA-10 v3 (PA-J v3) (N = 135)	PA-BT-10	short item	0.725	0.667
	PA-BT-11	short item		0.627
	<i>PA-EX-2a</i>	<i>extended item</i>		<i>0.700</i>
	<i>PA-EX-2b</i>	<i>extended item</i>		<i>0.657</i>
PA-11 v4 (PA-K v4) (N = 131)	PA-BT-3	short item	0.583	0.461
	PA-BT-7	short item		0.455
	<i>PA-EX-3a</i>	<i>extended item</i>		<i>0.652</i>
	<i>PA-EX-3b</i>	<i>extended item</i>		<i>0.534</i>
	<i>PA-EX-3c</i>	<i>extended item</i>		<i>0.505</i>
PA-12 v1 (N = 43)	PA-BT-10	short item	0.785	0.753
	PA-BT-11	short item		0.759
	PA-PW-6a	short item		0.764
	PA-PW-6b	short item		0.699
	PA-PW-6c	short item		0.747
PA-13 v2 (N = 84)	PA-BT-10	short item	0.526	0.303
	PA-BT-11	short item		0.380
	<i>PA-EX-6</i>	<i>extended item</i>		<i>0.569</i>
PA-14 v2 (N = 82)	PA-BT-10	short item	0.794	0.817
	PA-BT-11	short item		0.767
	PA-PW-7a	short item		0.726
	PA-PW-7b	short item		0.723
	PA-PW-7c	short item		0.728
PA-15 v1 (N = 28)	PA-BT-10	short item	0.643	0.609
	PA-BT-11	short item		0.643
	PA-PW-8a	short item		0.544
	PA-PW-8b	short item		0.627
	PA-PW-8c	short item		0.511



#### A4. SE: Year 1 Pilot Study

FORM	ITEM	Short/Extended	Reliability	Alpha if item deleted
SE-1 v3 (SE-A3) (N = 87 )	SE-BT-1	short item	0.470	0.531
	SE-BT-2	short item		0.429
	SE-PW-1-a	short item		0.340
	SE-PW-1-b	short item		0.361
	<i>SE-PWE-1</i>	<i>extended item</i>		<i>0.387</i>
SE-B v4 (SE-2-4) (SE-B7) (N = 80)	SE-BT-1	short item	0.561	0.604
	SE-BT-2	short item		0.473
	SE-PW-2-a	short item		0.424
	SE-PW-2-b	short item		0.603
	<i>SE-PWE-2</i>	<i>extended item</i>		<i>0.371</i>
SE-3 v1 (N = 17)	SE-BT-1	short item	-0.226	-0.402
	SE-BT-2	short item		0.118
	<i>SE-EX-1</i>	<i>extended item</i>		<i>-0.201</i>
SE-D v4 (SE-4-4) (N = 59)	SE-BT-3	short item	0.545	0.549
	SE-BT-4	short item		0.518
	SE-PW-1-a	short item		0.523
	SE-PW-1-b	short item		0.415
	<i>SE-PWE-1</i>	<i>extended item</i>		<i>0.426</i>
SE-5 v4 (N = 78)	SE-BT-5	short item	0.512	0.445
	SE-BT-6	short item		0.556
	SE-PW-1-a	short item		0.388
	SE-PW-1-b	short item		0.407
	<i>SE-PWE-1</i>	<i>extended item</i>		<i>0.467</i>
SE-7 v4 (N = 121)	SE-BT-1	short item	0.712	0.756
	SE-BT-5	short item		0.711
	<i>SE-WP-2-a</i>	<i>extended item</i>		<i>0.623</i>
	<i>SE-WP-2-b</i>	<i>extended item</i>		<i>0.622</i>
	<i>SE-WP-2-c</i>	<i>extended item</i>		<i>0.663</i>
	<i>SE-WP-2-d</i>	<i>extended item</i>		<i>0.668</i>
	<i>SE-WP-2-e</i>	<i>extended item</i>		<i>0.687</i>
SE-8 v1 (N = 109 )	SE-BT-5	short item	0.687	0.654
	SE-BT-6	short item		0.737
	SE-PW-3-a	short item		0.616
	SE-PW-3-b	short item		0.617
	SE-PW-3-c	short item		0.589
	<i>SE-EX-3</i>	<i>extended item</i>		<i>0.643</i>

**A5. SE: Year 2 Pilot Study (extended items are not scored)**

FORM	ITEM	Short/Extended	Reliability	Alpha if Item Deleted
SE-15 (N = 73)	SE-BT-7	short item	-0.266	
	SE-WP-7	extended item		
	SE-EX-10-a	short item		
	SE-EX-10-b	extended item		
	SE-EX-10-c	extended item		
SE-16 (N = 48)	SE-BT-8	short item	-0.216	
	SE-WP-8	extended item		
	SE-EX-11-a	short item		
	SE-EX-11-b	extended item		
	SE-EX-11-c	extended item		
SE-17 (N = 35)	SE-BT-9	short item	0.442	0.140
	SE-WP-9-a	short item		0.404
	SE-WP-9-b	extended item		
	SE-WP-9-c	extended item		
	SE-EX-12-a	short item		0.459
	SE-EX-12-b	extended item		
	SE-EX-12-c	extended item		
SE-18 (N = 18)	SE-BT-7	short item	0.064	0.264
	SE-WP-9-a	short item		-0.052
	SE-WP-9-b	extended item		
	SE-WP-9-c	extended item		
	SE-EX-13-a	short item		-0.127
	SE-EX-13-b	extended item		
	SE-EX-13-c	extended item		
SE-19 (N = 39)	SE-BT-8	short item	0.592	
	SE-WP-7	extended item		
	SE-EX-14-a	short item		
	SE-EX-14-b	extended item		
	SE-EX-14-c	extended item		

(table continues)

**A5. SE: Year 2 Pilot Study (extended items are not scored, continued))**

FORM	ITEM	Short/Extended	Reliability	Alpha if Item Deleted
SE-20 (N = 80)	SE-BT-9	short item	0.045	
	SE-WP-8	<i>extended item</i>		
	SE-EX-15-a	short item		
	SE-EX-15-b	<i>extended item</i>		
	SE-EX-15-c	<i>extended item</i>		
SE-21 (N = 36)	SE-BT-7	short item	0.072	
	SE-WP-8	<i>extended item</i>		
	SE-EX-16-a	short item		
	SE-EX-16-b	<i>extended item</i>		
	SE-EX-16-c	<i>extended item</i>		
SE-22 (N = 32)	SE-BT-8	short item	0.596	0.307
	SE-WP-9-a	short item		0.812
	SE-WP-9-b	<i>extended item</i>		
	SE-WP-9-c	<i>extended item</i>		
	SE-EX-10-a	short item		0.221
	SE-EX-10-b	<i>extended item</i>		
	SE-EX-10-c	<i>extended item</i>		
SE-23 (N = 75)	SE-BT-8	short item	0.462	
	SE-WP-7	<i>extended item</i>		
	SE-EX-11-a	short item		
	SE-EX-11-b	<i>extended item</i>		
	SE-EX-11-c	<i>extended item</i>		

## Appendix B

ITEM ID #	FORM 1	FORM 2	ITEM															
PA-BT-3	✓	✓	$6(3 + 1) = 6 \cdot \square + 6 \cdot 1$															
PA-BT-4	✓	✓	$3(15 + 5) = \square \cdot 15 + 3 \cdot 5$															
PA-PW-2abc		✓	<p>A student simplified the expression <math>4(5 + 2)</math> in 4 steps. Can you fill in the missing numbers in steps 2, 3 and 4?</p> <table><tr><th></th><th>Simplifying Step</th></tr><tr><td>1</td><td><math>4(5 + 2)</math></td></tr><tr><td>2</td><td><math>4 \cdot 5 + \square \cdot 2</math></td></tr><tr><td>3</td><td><math>20 + \square</math></td></tr><tr><td>4</td><td><math>\square</math></td></tr></table>		Simplifying Step	1	$4(5 + 2)$	2	$4 \cdot 5 + \square \cdot 2$	3	$20 + \square$	4	$\square$					
	Simplifying Step																	
1	$4(5 + 2)$																	
2	$4 \cdot 5 + \square \cdot 2$																	
3	$20 + \square$																	
4	$\square$																	
PA-PW-3ab	✓		<p>A student simplified the expression <math>2(7 + 4)</math> like this. Can you fill in the missing numbers in steps 2 and 3?</p> <table><tr><th></th><th>Simplifying Step</th></tr><tr><td>1</td><td><math>2(7 + 4)</math></td></tr><tr><td>2</td><td><math>2 \cdot 7 + 2 \cdot \square</math></td></tr><tr><td>3</td><td><math>14 + \square</math></td></tr><tr><td>4</td><td>22</td></tr></table>		Simplifying Step	1	$2(7 + 4)$	2	$2 \cdot 7 + 2 \cdot \square$	3	$14 + \square$	4	22					
	Simplifying Step																	
1	$2(7 + 4)$																	
2	$2 \cdot 7 + 2 \cdot \square$																	
3	$14 + \square$																	
4	22																	
PA-PWE-1	✓		<p>A student simplified the expression <math>5(3 + 2)</math> using 4 steps, but he forgot to write the explanations. Explain what the student did in step 2 and <i>why</i> he did it. Be sure to use some mathematical rule or principle in your explanation.</p> <table><tr><th></th><th>Simplifying Step</th><th>Explanation</th></tr><tr><td>1</td><td><math>5(3 + 2)</math></td><td>Wrote the problem.</td></tr><tr><td>2</td><td><math>5 \cdot 3 + 5 \cdot 2</math></td><td></td></tr><tr><td>3</td><td><math>15 + 10</math></td><td>Multiplied: <math>5 \cdot 3 = 15</math> and <math>5 \cdot 2 = 10</math></td></tr><tr><td>4</td><td>25</td><td>Added: <math>15 + 10</math></td></tr></table>		Simplifying Step	Explanation	1	$5(3 + 2)$	Wrote the problem.	2	$5 \cdot 3 + 5 \cdot 2$		3	$15 + 10$	Multiplied: $5 \cdot 3 = 15$ and $5 \cdot 2 = 10$	4	25	Added: $15 + 10$
	Simplifying Step	Explanation																
1	$5(3 + 2)$	Wrote the problem.																
2	$5 \cdot 3 + 5 \cdot 2$																	
3	$15 + 10$	Multiplied: $5 \cdot 3 = 15$ and $5 \cdot 2 = 10$																
4	25	Added: $15 + 10$																
PA-PWE-2		✓	<p>A student simplified the expression <math>4(12 + 3)</math> in 4 steps. Explain what the student did in step 2 and <i>why</i> he did it. Be sure to use some mathematical rule or principle in your explanation.</p> <table><tr><th></th><th>Simplifying Step</th><th>Explanation</th></tr><tr><td>1</td><td><math>4(12 + 3)</math></td><td>Wrote the problem.</td></tr><tr><td>2</td><td><math>4 \cdot 12 + 4 \cdot 3</math></td><td></td></tr><tr><td>3</td><td><math>48 + 12</math></td><td>Multiplied: <math>4 \cdot 12 = 48</math> and <math>4 \cdot 3 = 12</math></td></tr><tr><td>4</td><td>60</td><td>Added: <math>48 + 12</math></td></tr></table>		Simplifying Step	Explanation	1	$4(12 + 3)$	Wrote the problem.	2	$4 \cdot 12 + 4 \cdot 3$		3	$48 + 12$	Multiplied: $4 \cdot 12 = 48$ and $4 \cdot 3 = 12$	4	60	Added: $48 + 12$
	Simplifying Step	Explanation																
1	$4(12 + 3)$	Wrote the problem.																
2	$4 \cdot 12 + 4 \cdot 3$																	
3	$48 + 12$	Multiplied: $4 \cdot 12 = 48$ and $4 \cdot 3 = 12$																
4	60	Added: $48 + 12$																

## Appendix C

### IRT Tables

#### C1. RNE Years 1 and 2 Pilot

Item	<i>b</i>	<i>SE(b)</i>	tau1	tau2	tau3	tau4	G <sup>2</sup>	<i>df</i>	<i>p</i> -value	INFIT (MNSQ)	OUTFIT (MNSQ)
RNBT1	-1.52	0.09					11.72	5	0.04	1.07	3.46
RNBT2	-1.11	0.08					35.91	6	0.00	1.00	1.05
RNBT4	-0.65	0.08					10.38	6	0.11	1.17	1.75
RNBT5	-1.01	0.09					28.74	5	0.00	0.96	1.07
RNBT6	-0.72	0.07					38.96	7	0.00	1.04	1.34
RNBT7	-1.48	0.19					0.00	0	0.00	0.97	0.74
RNBT8	-0.25	0.08					25.33	8	0.00	1.11	1.10
RNBT10	-3.10	0.54					0.00	0	0.00	0.99	1.13
RNPW2A	0.70	0.12					10.36	6	0.11	1.00	1.36
RNPW2B	0.78	0.12					27.31	7	0.00	0.79	0.62
RNPW2C	0.81	0.13					26.49	7	0.00	0.86	0.71
RNPW3A	-0.41	0.16					1.73	2	0.42	0.87	0.96
RNPW3B	-1.06	0.19					0.00	0	0.00	0.96	0.84
RNPW3C	-0.50	0.16					4.07	2	0.13	0.83	0.80
RNPW4A	0.02	0.15					9.05	4	0.06	0.82	0.62
RNPW4B	-1.07	0.16					0.20	2	0.90	1.06	0.86
RNPW4C	-0.48	0.15					6.80	3	0.08	0.74	0.67
RNEEX1_A	-0.61	0.07	-0.53	-0.01	0.55		10.06	3	0.02	0.61	0.82
RNEEX1_B	1.06	0.08	1.39	0.34	-1.73		92.64	11	0.00	0.82	0.67
RNEEX2_A	0.24	0.08	-0.17	0.17			18.87	7	0.01	1.18	1.19
RNEEX2_B	0.05	0.05	-0.37	0.11	0.06	0.20	8.72	6	0.19	0.49	0.53
RNEEX3_A	0.50	0.07	-0.05	0.90	-0.85		16.85	12	0.16	1.04	1.00
RNEEX3_B	0.06	0.08	-0.26	0.26			21.46	6	0.00	0.76	0.80
RNEEX3_C	1.42	0.09	0.63	0.50	-1.14		19.27	9	0.02	0.96	0.71
RNEEX6_A	0.74	0.12					20.13	4	0.00	0.91	2.81
RNEEX6_B	1.29	0.11	-0.47	0.47			18.68	6	0.01	0.86	6.02
RNEEX6_C	0.94	0.13					9.27	5	0.10	1.25	2.12
RNEEX6_D	1.61	0.15	-0.74	0.74			11.00	6	0.09	0.89	0.15
RNEPWE1_	0.94	0.16	0.44	-0.44			3.45	4	0.49	1.03	1.02
RNEPWE_1	1.02	0.15	0.23	-0.77	0.55		3.54	4	0.47	1.11	0.86
RNEWP2_A	0.71	0.17					10.77	4	0.03	1.17	4.61
RNEWP2_B	1.48	0.13	0.02	0.70	-0.71		8.93	5	0.11	0.90	0.23

## C2. RNE Field Test

Item	$b$	$SE(b)$	tau1	tau2	tau3	$G^2$	$df$	$p$ -value	INFIT (MNSQ)	OUTFIT (MNSQ)
RNBT7	-2.48	0.08				8.50	5	0.13	1.19	2.15
RNBT8	-0.57	0.02				258.57	10	0.00	0.88	0.79
RNBT9	-0.50	0.02				275.73	10	0.00	0.87	0.78
RNBT4	-0.83	0.03				85.63	9	0.00	0.99	0.94
RNBT2	-0.83	0.03				157.39	9	0.00	0.94	0.83
EX8	0.76	0.02	0.24	-1.03	0.79	113.72	25	0.00	1.05	1.22
EX9	-0.22	0.01	-1.26	0.45	0.81	96.68	23	0.00	1.24	2.10
EX1A	-0.11	0.01	-0.66	-0.02	0.68	117.15	26	0.00	0.94	0.98
EX1B	0.97	0.02	0.60	-0.19	-0.41	330.23	27	0.00	0.89	0.79
PWE_2	0.81	0.02	-0.20	0.44	-0.24	32.37	25	0.15	1.03	1.22
EX6a1	0.15	0.04				44.15	10	0.00	1.12	1.28
EX6a2	1.08	0.03	-0.50	0.02	0.48	25.07	20	0.20	1.08	1.50
EX6b1	0.70	0.04				66.66	10	0.00	1.12	1.75
EX6b2	1.26	0.04	-0.76	0.37	0.39	31.88	16	0.01	0.93	1.23

### C3. PA Year 1 Pilot

Item	<i>b</i>	<i>SE(b)</i>	tau1	tau2	tau3	$G^2$	<i>df</i>	<i>p</i> -value	INFIT (MNSQ)	OUTFIT (MNSQ)
pabt3	-0.85	0.07				9.67	4	0.05	0.99	1.41
pabt4	0.10	0.08				10.93	3	0.01	1.19	2.35
pabt5	-0.63	0.29				0.00	0	0.00	0.96	0.82
pabt6	0.29	0.28				0.40	2	0.82	1.56	1.59
pabt7	-0.37	0.10				14.01	3	0.00	0.82	0.77
pabt8	-0.34	0.19				1.28	2	0.53	0.97	0.74
pabt9	0.27	0.19				0.92	2	0.64	1.24	1.56
pabt10	-0.65	0.08				23.11	4	0.00	1.05	1.16
pabt11	-0.27	0.07				33.46	4	0.00	0.98	1.1
papw2a	0.69	0.14				7.52	3	0.06	0.68	0.46
papw2b	0.66	0.14				13.91	3	0.00	0.57	0.38
papw2c	0.80	0.15				15.10	3	0.00	0.6	0.37
papw3a	-0.12	0.18				0.74	1	0.39	0.8	0.7
papw3b	-0.41	0.19				0.95	1	0.33	0.84	0.59
papw4a	-0.35	0.16				10.78	2	0.01	0.7	0.51
papw4b	-0.01	0.16				9.95	2	0.01	0.74	0.61
papw5a	0.29	0.18				0.41	2	0.82	1.25	1.42
papw6a	0.91	0.28				0.02	2	0.98	1.27	1.54
papw6b	0.56	0.25				4.45	2	0.11	0.74	0.59
papw6c	0.67	0.26				0.37	2	0.83	1.08	1.11
papw7a	-0.06	0.16				17.09	4	0.00	0.84	0.73
papw7b	0.15	0.16				18.32	4	0.00	0.83	0.66
papw7c	-0.29	0.16				19.90	4	0.00	0.76	0.7
papw8a	-0.04	0.26				1.03	2	0.60	0.82	0.79
papw8b	0.46	0.28				0.03	2	0.98	1.13	1.26
papw8c	0.91	0.32				1.01	2	0.61	0.77	0.49
paex1_a	1.59	0.14	0.59	0.19	-0.79	1.85	4	0.77	0.82	0.73
paex1_b	0.50	0.09	-0.69	-0.73	1.42	0.83	2	0.66	0.47	0.96
paex2_a	1.20	0.10	0.13	-0.13		15.62	5	0.01	1.2	1.12
paex2_b	0.59	0.06	-0.13	-1.08	1.21	5.04	4	0.28	0.68	0.94
paex3_a	-1.60	0.18				2.40	1	0.12	1.36	1.58
paex3_b	1.18	0.15				2.85	3	0.42	1.05	0.88
paex3_c	1.32	0.10	1.34	-0.35	-0.99	12.34	7	0.09	1.08	0.99
paex6	1.15	0.11	0.62	-1.13	0.51	4.43	4	0.35	1.08	1.46
papwe1	1.67	0.18	-0.82	0.82		23.77	3	0.00	1.22	9.9
papwe2	2.06	0.19	-0.21	0.21		15.42	4	0.00	1.41	5.89
papwe3	1.81	0.13	-0.25	0.96	-0.71	6.09	4	0.19	0.79	3.09

#### C4. PA Field

Item	<i>b</i>	<i>SE</i> ( <i>b</i> )	tau1	tau2	tau3	$G^2$	<i>df</i>	<i>p</i> -value	INFIT (MNSQ)	OUTFIT (MNSQ)
PABT1_1R	-0.25	0.02				51.04	10	0	1.09	1.22
PABT13_1	-0.05	0.02				81.43	10	0	0.96	0.99
PABT9_1R	-0.39	0.02				33.23	9	0	1.08	1.1
PAPW8A_1	0.32	0.02				108.31	10	0	1.07	1.31
PAPW8B_1	0.57	0.02				195.87	10	0	0.85	1.02
PAPW8C_1	0.56	0.02				163.49	10	0	0.89	1.25
PABT2_2R	-0.98	0.02				46.02	8	0	0.93	0.94
PABT3_2R	-1.15	0.02				72.56	8	0	0.86	0.75
PABT12_2	-1.00	0.02				49.50	8	0	0.93	0.78
PAPW4A_2	-1.21	0.02				89.94	7	0	0.85	0.58
PAPW4B_2	-0.67	0.02				233.04	9	0	0.71	0.58
PAPW4C_2	-0.58	0.02				184.22	9	0	0.76	0.75
PAEX1a	1.61	0.03	-0.71	0.71		85.05	13	0	1.22	5.3
PAEX1b	0.34	0.02	-0.36	0.36		25.33	13	0.02	1.14	1.6
PAPWE3	1.84	0.02	0.67	0.80	-1.47	148.80	20	0	1.46	2.64
PAEX6a	1.59	0.02	0.77	0.65	-1.42	39.04	21	0.01	1.12	1.45
PAEX6b	-0.37	0.03				137.70	8	0	1.58	2.01



### C5. SE Years 1 and 2 Pilot

Item	<i>b</i>	<i>SE(b)</i>	tau1	tau2	tau3	G <sup>2</sup>	<i>df</i>	<i>p</i> -value	INFIT (MNSQ)	OUTFIT (MNSQ)
sebt1	-2.12	0.18				0.02	1	0.87	1.31	1.49
sebt2	-0.31	0.10				0.38	2	0.83	1.05	0.98
sebt3	-1.80	0.33				0.00	0	0.00	1.10	3.25
sebt4	-2.01	0.37				0.00	0	0.00	0.92	0.71
sebt5	0.13	0.08				2.74	4	0.61	1.25	1.63
sebt6	-1.75	0.17				1.55	1	0.21	1.27	1.80
sebt7	-1.15	0.15				10.49	2	0.01	1.05	2.23
sebt8	-1.38	0.15				14.37	2	0.00	0.78	0.56
sebt9	-0.33	0.15				4.08	2	0.13	0.93	0.88
sepw1a	-0.81	0.10				19.29	2	0.00	0.80	0.80
sepw1b	-0.77	0.10				20.70	2	0.00	0.73	0.64
sepw2a	-0.01	0.15				3.44	2	0.18	0.95	1.13
sepw2b	-0.72	0.17				2.01	2	0.37	1.25	1.26
sepw3a	-0.29	0.13				1.26	2	0.54	0.90	0.78
sepw3b	0.44	0.14				2.70	3	0.44	1.04	0.92
sepw3c	-0.44	0.13				4.27	2	0.12	0.76	0.61
seex1	0.49	0.25	1.67	-0.63	-1.04	0.59	1	0.45	0.85	0.80
seex3	0.36	0.07	0.07	0.52	-0.59	2.57	5	0.77	1.12	0.93
seex10a	-1.09	0.17				9.78	2	0.01	1.20	0.87
seex11a	-1.62	0.20				0.00	0	0.00	0.94	0.61
seex12a	-2.03	0.45				0.00	0	0.00	0.94	0.49
seex13a	-0.49	0.33				0.01	1	0.87	1.34	1.79
seex14a	-1.09	0.27				0.00	0	0.00	0.99	0.98
seex15a	-1.56	0.22				0.00	1	0.91	1.13	1.02
seex16a	-1.85	0.38				0.00	0	0.00	1.00	1.00
sepwe1	0.69	0.10				20.05	3	0.00	1.06	1.21
sepwe2	0.78	0.12	0.09	-0.09		6.51	4	0.16	0.79	0.70
sewp2a	-0.93	0.13	0.93	-0.93		2.50	2	0.29	0.55	0.45
sewp2b	-0.91	0.13	0.95	-0.95		2.92	2	0.23	0.56	0.46
sewp2c	1.38	0.20				2.75	3	0.43	0.94	0.74
sewp2d	1.53	0.22				1.41	3	0.71	1.34	9.90
sewp2e	0.07	0.11	0.65	-0.65		2.36	5	0.80	1.21	1.12
sewp9a	1.25	0.20				23.12	3	0.00	1.01	0.70

## C6. SE Field

Item	$b$	$SE(b)$	tau1	tau2	tau3	$G^2$	$df$	$p$ -value	INFIT (MNSQ)	OUTFIT (MNSQ)
sebt2	-0.21	0.03				246.52	10	0.00	0.94	0.92
sebt3	-3.35	0.14				0	0	0.00	1.11	2.44
sebt7	-1.35	0.03				28.5557	9	0.00	1.14	1.91
sebt8	-1.56	0.04				51.6484	8	0.00	1.08	1.90
sepw2a	0.01	0.02				162.165	10	0.00	1.12	1.31
sepw2b	-0.67	0.03				161.148	10	0.00	0.99	1.03
sepw3a	-0.38	0.03				352.773	10	0.00	0.85	0.83
sepw3b	0.33	0.03				262.473	10	0.00	0.99	1.11
sepw3c	-0.46	0.03				382.581	10	0.00	0.83	0.87
seex6a	-0.12	0.03	0.90	-0.90		206.559	17	0.00	0.83	0.82
seex6b	1.52	0.03	0.44	0.10	-0.55	156.45	20	0.00	0.80	0.60
seex9	1.14	0.03	0.37	-0.23	-0.14	53.9213	24	0.00	1.07	1.36
sepwe2	1.58	0.04	0.46	-0.12	-0.34	45.2868	19	0.00	1.08	1.02

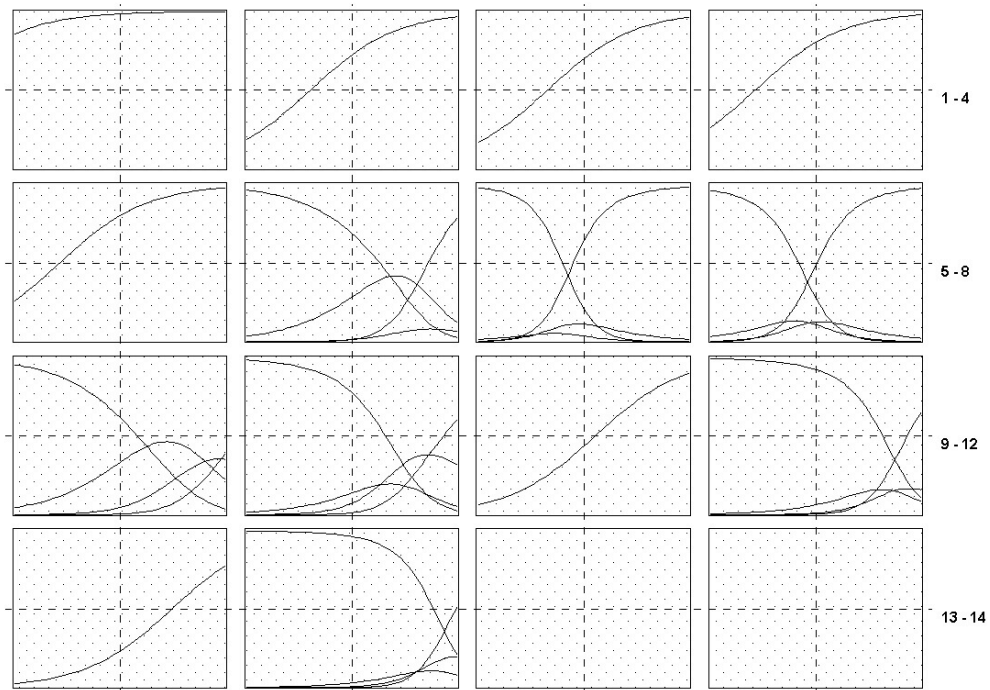
## C7. RA Field

Item	<i>b</i>	<i>SE(b)</i>	tau1	tau2	tau3	$G^2$	<i>df</i>	<i>p</i> -value	INFIT (MNSQ)	OUTFIT (MNSQ)
rasebt9	-1.03	0.05				41.28	7	0.00	1.30	1.98
rasebt10	-1.65	0.06				6.00	5	0.31	1.00	1.53
rasebt11	-0.41	0.04				15.98	8	0.04	0.97	1.03
rpabt14a	-0.66	0.04				46.30	8	0.00	0.84	0.69
rpabt14b	-0.64	0.04				20.95	8	0.01	1.02	0.95
rpabt14c	-0.24	0.04				78.73	9	0.00	0.78	0.68
rpabt15a	-0.58	0.04				25.15	8	0.00	0.90	0.75
rpabt15b	-0.42	0.04				26.25	8	0.00	1.10	1.15
rarnbt11	-1.62	0.06				26.07	5	0.00	0.86	0.87
rarnbt13	-1.40	0.05				30.73	7	0.00	0.83	0.72
rarnwp3	-0.41	0.02	-0.55	0.49	0.06	60.81	17	0.00	1.25	1.61
rarnwp4	0.75	0.05				44.99	10	0.00	1.04	0.99
rarngp1	-0.80	0.05				5.68	8	0.68	1.01	1.16
rarnbt14	0.01	0.04				22.24	9	0.01	1.01	0.96
rapabt16	0.14	0.07				43.56	8	0.00	0.69	0.6
rasewp10	0.63	0.04	0.61	-0.60	-0.01	16.08	15	0.38	1.10	1.31
rasegp1	-0.08	0.06	1.19	-1.19		21.43	14	0.09	1.23	1.27
rapabt18	0.55	0.05	0.21	-0.21		18.76	14	0.17	0.95	0.93
rapabt17	0.28	0.07				49.03	9	0.00	0.78	0.79

## Appendix D

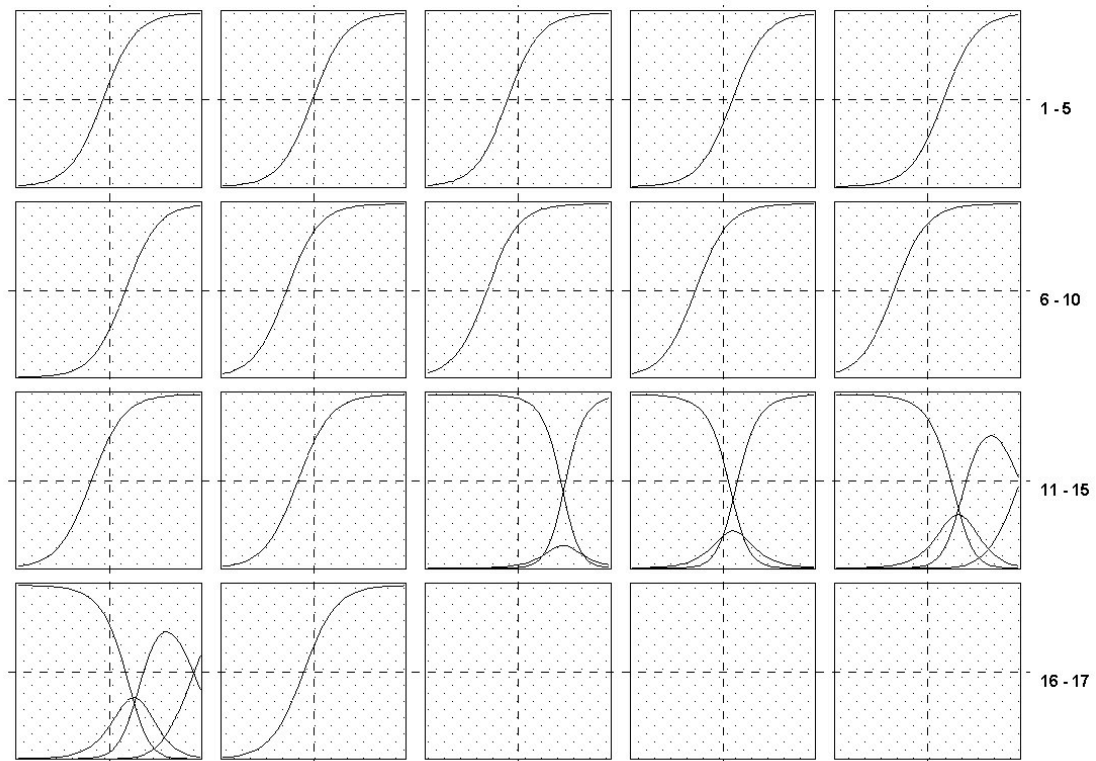
### Item (Category) Characteristic Curves

#### D1. Items in RNE Field Test



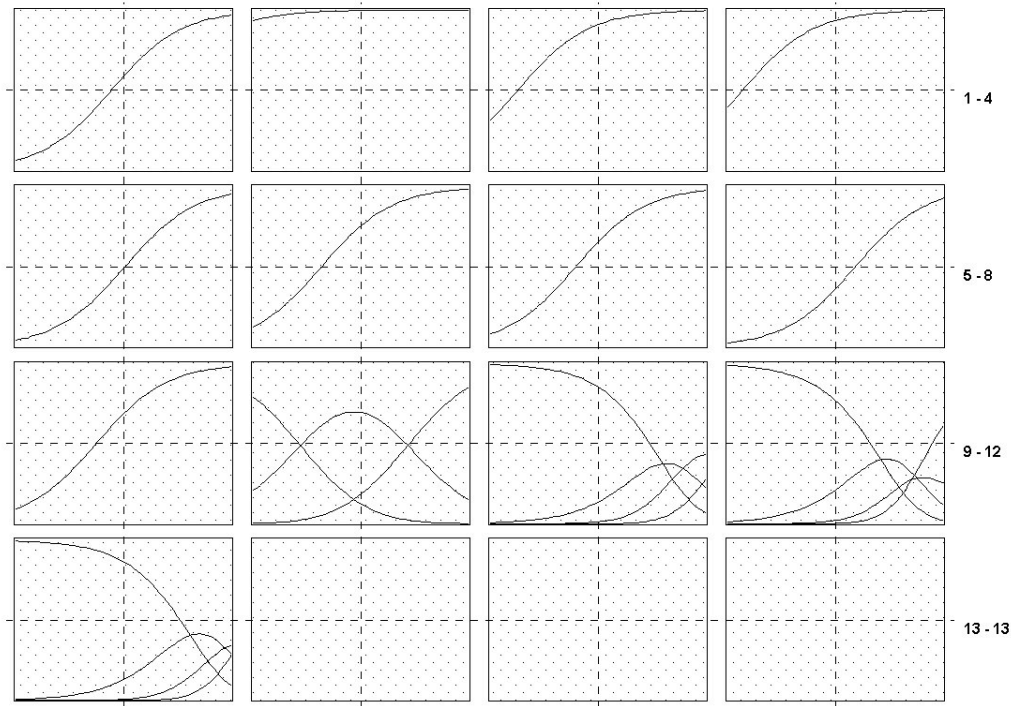
RNBT7	RNBT8	RNBT9	RNBT4
RNBT2	EX8	EX9	EX1A
EX1B	PWE_2	EX6a1	EX6a2
EX6b1	EX6b2		

## D2. Items in PA Field Test



PABT1_1R	PABT13_1	PABT9_1R	PAPW8A_1	PAPW8B_1
PAPW8C_1	PABT2_2R	PABT3_2R	PABT12_2	PAPW4A_2
PAPW4B_2	PAPW4C_2	PAEX1a	PAEX1b	PAPWE3
PAEX6a	PAEX6b			

**D3. Items in SE Field Test**



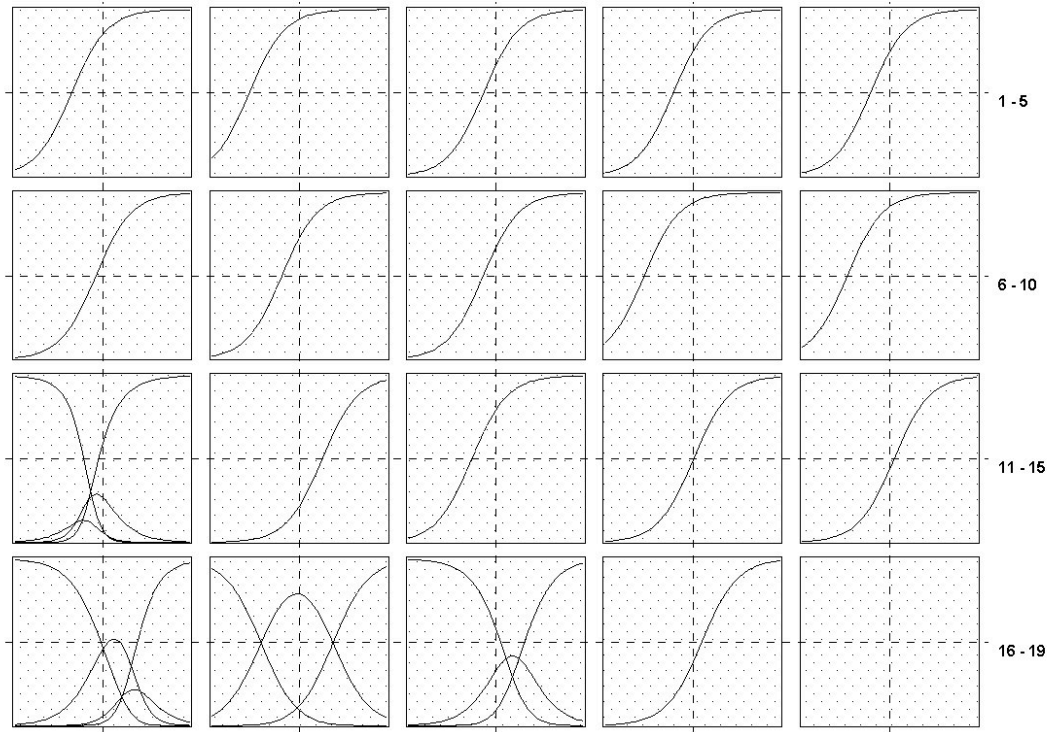
sebt2      sebt3      sebt7      sebt8

sepw2a    sepw2b    sepw3a    sepw3b



sepw3c    seex6a    seex6b    seex9

sepwe2

#### D4. Items in RA Field Test



## Appendix E

ITEM ID #	FORM 18	FORM 19	ITEM										
PA-BT-1	✓		$6(3 + 1) = 6 \square 3 + 6 \cdot 1$										
PA-BT-2		✓	$3(15 + 5) = 3 \cdot 15 \square 3 \cdot 5$										
PA-BT-3		✓	$6(3 + 1) = 6 \cdot \square + 6 \cdot 1$										
PA-BT-9	✓		$3(15 + 5) = 3 \cdot \square + 3 \cdot 5$										
PA-BT-12		✓	$3(15 + 5) = (\square \cdot 15) + (3 \cdot 5)$										
PA-BT-13	✓		$7(4 + 2) = \square \cdot 4 + 7 \cdot 2$										
PA-PW-4abc		✓	<p>A student simplified the expression <math>2(7 + 4)</math> like this. Can you fill in the missing numbers in steps 2,3 and 4?</p> <table><tr><th></th><th>Simplifying Step</th></tr><tr><td>1</td><td><math>2(7 + 4)</math></td></tr><tr><td>2</td><td><math>(2 \times 7) + (2 \times \square)</math></td></tr><tr><td>3</td><td><math>14 + \square</math></td></tr><tr><td>4</td><td><math>\square</math></td></tr></table>		Simplifying Step	1	$2(7 + 4)$	2	$(2 \times 7) + (2 \times \square)$	3	$14 + \square$	4	$\square$
	Simplifying Step												
1	$2(7 + 4)$												
2	$(2 \times 7) + (2 \times \square)$												
3	$14 + \square$												
4	$\square$												
PA-PW-8abc	✓		<p>A student used the distributive property to show the total area of both the rectangles below. Can you fill in the missing numbers for steps 1, 2, and 3?</p> <div> </div> <table><tr><td>1</td><td><math>\square(3 + \square) = 20</math></td></tr><tr><td>2</td><td><math>(\square \cdot 3) + (\square \cdot \square) = 20</math></td></tr><tr><td>3</td><td><math>\square + \square = 20</math></td></tr></table>	1	$\square(3 + \square) = 20$	2	$(\square \cdot 3) + (\square \cdot \square) = 20$	3	$\square + \square = 20$				
1	$\square(3 + \square) = 20$												
2	$(\square \cdot 3) + (\square \cdot \square) = 20$												
3	$\square + \square = 20$												



cont. Items in Test Forms PA-18 and PA-19

PA-PWE-3

A student simplified the expression  $3(y + 3)$  in 5 steps. Explain what the student did in step 2 and why he did it. Be sure to use some mathematical rule or principle in your explanation.

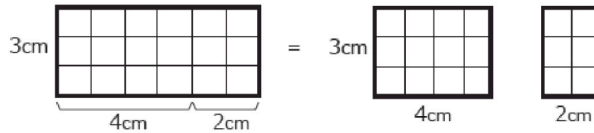
	Simplifying Step	Explanation
1	$3(y + 3)$	Wrote the problem.
2	$(y + 3) + (y + 3) + (y + 3)$	
3	$y + y + y + 3 + 3 + 3$	Addition is commutative.
4	$(3 \times y) + (3 \times 3)$	Combined like terms using multiplication.
5	$3y + 9$	Multiplied $3 \times y$ to get $3y$ and $3 \times 3$ to get 9.

PA-EX-1ab

In the space below, explain "the distributive property" and give an example of how to use it.

PA-EX-6ab

A student drew these diagrams to show how the distributive property works. Do you agree that these diagrams show how the distributive property works? Explain your answer.



cont. Items in Test Forms PA-18 and PA-19

PA-PWE-3

A student simplified the expression  $3(y + 3)$  in 5 steps. Explain what the student did in step 2 and why he did it. Be sure to use some mathematical rule or principle in your explanation.

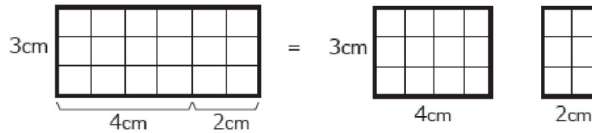
	Simplifying Step	Explanation
1	$3(y + 3)$	Wrote the problem.
2	$(y + 3) + (y + 3) + (y + 3)$	<div style="border: 1px solid black; height: 50px; width: 100%;"></div>
3	$y + y + y + 3 + 3 + 3$	Addition is commutative.
4	$(3 \times y) + (3 \times 3)$	Combined like terms using multiplication.
5	$3y + 9$	Multiplied $3 \times y$ to get $3y$ and $3 \times 3$ to get 9.

PA-EX-1ab

In the space below, explain "the distributive property" and give an example of how to use it.

PA-EX-6ab

A student drew these diagrams to show how the distributive property works. Do you agree that these diagrams show how the distributive property works? Explain your answer.



ITEM ID #	FORM 18	FORM 19	ITEM																		
PA-PWE-3		✓	<p>A student simplified the expression <math>3(y + 3)</math> in 5 steps. Explain what the student did in step 2 and why he did it. Be sure to use some mathematical rule or principle in your explanation.</p> <table><tr><th></th><th>Simplifying Step</th><th>Explanation</th></tr><tr><td>1</td><td><math>3(y + 3)</math></td><td>Wrote the problem.</td></tr><tr><td>2</td><td><math>(y + 3) + (y + 3) + (y + 3)</math></td><td></td></tr><tr><td>3</td><td><math>y + y + y + 3 + 3 + 3</math></td><td>Addition is commutative.</td></tr><tr><td>4</td><td><math>(3 \times y) + (3 \times 3)</math></td><td>Combined like terms using multiplication.</td></tr><tr><td>5</td><td><math>3y + 9</math></td><td>Multiplied <math>3 \times y</math> to get <math>3y</math> and <math>3 \times 3</math> to get 9.</td></tr></table>		Simplifying Step	Explanation	1	$3(y + 3)$	Wrote the problem.	2	$(y + 3) + (y + 3) + (y + 3)$		3	$y + y + y + 3 + 3 + 3$	Addition is commutative.	4	$(3 \times y) + (3 \times 3)$	Combined like terms using multiplication.	5	$3y + 9$	Multiplied $3 \times y$ to get $3y$ and $3 \times 3$ to get 9.
	Simplifying Step	Explanation																			
1	$3(y + 3)$	Wrote the problem.																			
2	$(y + 3) + (y + 3) + (y + 3)$																				
3	$y + y + y + 3 + 3 + 3$	Addition is commutative.																			
4	$(3 \times y) + (3 \times 3)$	Combined like terms using multiplication.																			
5	$3y + 9$	Multiplied $3 \times y$ to get $3y$ and $3 \times 3$ to get 9.																			
PA-EX-1ab	✓		<p>In the space below, explain “the distributive property” and give an example of how to use it.</p>																		
PA-EX-6ab		✓	<p>A student drew these diagrams to show how the distributive property works. Do you agree that these diagrams show how the distributive property works? Explain your answer.</p> <div><div><div>3cm</div><div></div></div><div>=</div><div><div><div>3cm</div><div></div></div><div><div></div></div></div></div>																		

## Appendix F

Reliability and factor analysis for RNE *Checks for Understanding* forms including in field testing

Test form			Reliability
Items	Short/Extended	Component loading	Alpha if item deleted
<b>RNE-10 (n = 3,320)</b>			<b>0.668</b>
RN-BT-7	short item	0.217	0.700
RN-BT-8	short item	0.779	0.587
RN-BT-9	short item	0.784	0.584
<i>RN-EX-8</i>	<i>extended item</i>	<i>0.407</i>	<i>0.666</i>
<i>RN-EX-9</i>	<i>extended item</i>	<i>0.554</i>	<i>0.640</i>
<i>RN-EX-1a</i>	<i>extended item</i>	<i>0.648</i>	<i>0.609</i>
<i>RN-EX-1b</i>	<i>extended item</i>	<i>0.581</i>	<i>0.627</i>
<b>RNE-9 (n = 3,068)</b>			<b>0.554</b>
RN-BT-4	short item	0.473	0.522
RN-BT-2	short item	0.461	0.526
<i>RN-PWE-2</i>	<i>extended item</i>	<i>0.514</i>	<i>0.520</i>
<i>RN-EX-6a</i>	<i>extended item</i>	<i>0.516</i>	<i>0.517</i>
<i>RN-EX-6b</i>	<i>extended item</i>	<i>0.595</i>	<i>0.500</i>
<i>RN-EX-6c</i>	<i>extended item</i>	<i>0.463</i>	<i>0.540</i>
<i>RN-EX-6d</i>	<i>extended item</i>	<i>0.626</i>	<i>0.486</i>

Reliability and factor analysis for SE *Checks for Understanding* forms including in field testing

Test form			Reliability
Items	Short/Extended	Component loading	Alpha if item deleted
<b>SE-11 v2 (n = 2,978)</b>			<b>0.641</b>
SE-BT-7	short item	0.377	0.645
SE-BT-8	short item	0.344	0.650
SE-PW-3a	short item	0.802	0.536
SE-PW-3b	short item	0.688	0.573
SE-PW-3c	short item	0.801	0.535
<i>SE-EX-9</i>	<i>extended item</i>	<i>0.479</i>	<i>0.627</i>
<b>SE-12 v1 (n = 2,961)</b>			<b>0.718</b>
SE-BT-3	short item	0.245	0.752
SE-BT-2	short item	0.635	0.677
SE-PW-2a	short item	0.624	0.685
SE-PW-2b	short item	0.531	0.701
<i>SE-PWE-2</i>	<i>extended item</i>	<i>0.636</i>	<i>0.684</i>
<i>SE-EX-6a</i>	<i>extended item</i>	<i>0.806</i>	<i>0.629</i>
<i>SE-EX-6b</i>	<i>extended item</i>	<i>0.751</i>	<i>0.655</i>

Reliability and factor analysis for RA *Checks for Understanding* forms including in field testing

Test form		Reliability	
Items	Short/Extended	Component loading	Alpha if item deleted
<b>RA-1 v3 (<i>n</i> = 1,163)</b>			<b>0.863</b>
RA-SE-BT-9	short item	0.374	0.867
RA-SE-BT-10	short item	0.463	0.861
RA-SE-BT-11	short item	0.553	0.854
RA-PA-BT-14a	short item	0.721	0.845
RA-PA-BT-14b	short item	0.645	0.852
RA-PA-BT-14c	short item	0.763	0.843
RA-PA-BT-15a	short item	0.739	0.847
RA-PA-BT-15b	short item	0.616	0.854
RA-RN-BT-11	short item	0.554	0.856
RA-RN-BT-13	short item	0.608	0.852
RA-RN-WP-3	short item (*range: 0–3)	0.696	0.848
<i>RA-PA-BT-16</i>	<i>extended item</i>	<i>0.739</i>	<i>0.846</i>
<b>RA-2 v2 (<i>n</i> = 1,111)</b>			<b>0.831</b>
RA-RN-WP-4	short item	0.703	0.811
RA-RN-GP-1	short item	0.558	0.830
RA-RN-BT-14	short item	0.725	0.806
<i>RA-SE-GP-1</i>	<i>extended item</i>	<i>0.674</i>	<i>0.814</i>
<i>RA-SE-WP-10</i>	<i>extended item</i>	<i>0.746</i>	<i>0.801</i>
<i>RA-PA-BT-17</i>	<i>extended item</i>	<i>0.728</i>	<i>0.804</i>
<i>RA-PA-BT-18</i>	<i>extended item</i>	<i>0.801</i>	<i>0.791</i>